AO

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68, C12M 3/00, C12N 15/00 — A1

(11) International Publication Number: WO 96/27025

(43) International Publication Date: 6 September 1996 (06.09.96)

(21) International Application Number: PCT/US96/02342

(22) International Filing Date: 21 February 1996 (21.02.96)

(30) Priority Data:
08/394,307 27 February 1995 (27.02.95) US

(71)(72) Applicant and Inventor: RABANI, Ely, Michael [US/US]; 4495 Vision Drive #1, San Diego, CA 92121-1942 (US).

(74) Agent: COTA, Albert, O.; Suite A-331, 5460 White Oak Avenue, Encino, CA 91316 (US).

(81) Designated States: AM, AT, AU, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, EE, ES, FI, GB, GE, HU, IS, JP, KE, KG, KP, KR, KZ, LK, LR, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, TJ, TT, UA, UG, UZ, VN, ARIPO patent (KE, LS, MW, SD, SZ, UG), European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: DEVICE, COMPOUNDS, ALGORITHMS, AND METHODS OF MOLECULAR CHARACTERIZATION AND MANIPULATION WITH MOLECULAR PARALLELISM

(57) Abstract

Methods and means are provided for the massively parallel characterization of complex molecules and of molecular recognition phenomena with parallelism and redundancy attained through single molecule examination methods. Applications include ultra-rapid genome sequencing, affinity characterization, pathogen characterization and detection means for clinical use and use in the development and construction of cybernetic immune systems. Novel methods for single molecule examination and manipulation are provided, including scanned beam light microscopic means and methods, and detection means availing of optoelectronic array devices. Various apparatus for rate control, including stepping control for various reactions are combined with molecular recognition, signal amplification and single molecule examination methods. Inclusion of internal control in samples, algorithm-based dynamically responsive manipulation controls, and sample redundancy, are availed to provide an arbitrarily high degree of accuracy in final data.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	**GB**	United Kingdom	**MW**	Malawi
AT	Austria	**GE**	Georgia	**MX**	Mexico
AU	Australia	**GN**	Guinea	**NE**	Niger
BB	Barbados	**GR**	Greece	**NL**	Netherlands
BE	Belgium	**HU**	Hungary	**NO**	Norway
BF	Burkina Faso	**IE**	Ireland	**NZ**	New Zealand
BG	Bulgaria	**IT**	Italy	**PL**	Poland
BJ	Benin	**JP**	Japan	**PT**	Portugal
BR	Brazil	**KE**	Kenya	**RO**	Romania
BY	Belarus	**KG**	Kyrgystan	**RU**	Russian Federation
CA	Canada	**KP**	Democratic People's Republic of Korea	**SD**	Sudan
CF	Central African Republic			**SE**	Sweden
CG	Congo	**KR**	Republic of Korea	**SG**	Singapore
CH	Switzerland	**KZ**	Kazakhstan	**SI**	Slovenia
CI	Côte d'Ivoire	**LI**	Liechtenstein	**SK**	Slovakia
CM	Cameroon	**LK**	Sri Lanka	**SN**	Senegal
CN	China	**LR**	Liberia	**SZ**	Swaziland
CS	Czechoslovakia	**LT**	Lithuania	**TD**	Chad
CZ	Czech Republic	**LU**	Luxembourg	**TG**	Togo
DE	Germany	**LV**	Latvia	**TJ**	Tajikistan
DK	Denmark	**MC**	Monaco	**TT**	Trinidad and Tobago
EE	Estonia	**MD**	Republic of Moldova	**UA**	Ukraine
ES	Spain	**MG**	Madagascar	**UG**	Uganda
FI	Finland	**ML**	Mali	**US**	United States of America
FR	France	**MN**	Mongolia	**UZ**	Uzbekistan
GA	Gabon	**MR**	Mauritania	**VN**	Viet Nam

DEVICE, COMPOUNDS, ALGORITHMS, AND METHODS OF MOLECULAR CHARACTERIZATION AND MANIPULATION WITH MOLECULAR PARALLELISM

Field of the Invention:

The invention relates to the massively parallel single molecule examination of associations or reactions between large numbers of first complex molecules, which may be diverse, and second single or plural probing molecules, which may or may not be diverse, with applications to biology, biotechnology, pharmacology, immunology, the novel field of cybernetic immunology, molecular evolution, cybernetic molecular evolution, genomics, comparative genomics, enzymology, clinical enzymology, pathology, medical research, and clinical medicine.

Background of the Invention:

Prior art polynucleotide sequence determination and characterization methods:

The present invention has applications in the area of polynucleotide sequence determination, including DNA sequencing.

Presently, there exist only two well established general methods for the determination of the presence, at a particular location within a sample polynucleotide molecule, of a particular base. These methods are: (1.) Sanger enzymatic chain-termination sequencing, which relies on the template directed incorporation of nucleotides which themselves do not supply the necessary chemical functionalities required for subsequent enzymatic polymerization of a daughter strand polynucleotide; and (2.) Maxam and Gilbert base-specific chemical modification and cleavage, which similarly yields polynucleotide molecules terminated at sites containing a specific base according to the chemical treatment applied to the sample. No distinct method preferable to these has yet been validated.

Both of these methods yield a population of molecules comprising a nested set which together may be analyzed to determine the base sequence of the sample. Analysis methods have heretofore relied on electrophoretic separation and resolution of the products of Sanger or Maxam and Gilbert reactions according to the length of said products. Analysis thus suffers all of the limitations associated with electrophoresis including limited separation range (i.e. limited dynamic range, where separative resolution is related exponentially to fractional differences in molecular length), limitations on parallelism, time requirements, etc., despite much effort in improving these separative methodologies.

These methods and several variations thereupon, as well as their severe limitations with respect to the economy and rapidity of accumulation of

sequence data, are well known to those in the relevant arts. Various lower resolution techniques, generally falling within the category termed genome mapping, have been developed to circumvent these limitations for applications where more "broad spectrum" examination of genetic material is required but less detailed information about sequence will suffice. Mapping techniques include restriction enzyme analysis of genetic material, and the hybridization and detection of specific oligonucleotides which test for the presence or absence of particular alleles or loci, and may further be used to gain spatial information about the occurrence of their targets when appropriate analytic techniques are subsequently applied. Note that such characterizations presently are methodologically and operationally distinct from other processes comprehended within the biotechnological and related arts.

Prior art 3'-hydroxy-protected and labeled nucleotides:

A modified nucleotide compound possessing two properties particularly useful for purposes of the present invention has been described by N. Williams and P.S. Coleman[1]. This compound is 3'-O-(4-benzoyl)benzoyl adenosine 5'-triphosphate. This nucleotide bears a 3' protecting group linked via an ester function which should be susceptible to hydrolysis by appropriate chemical treatments. The protecting moiety is suitable for photoactivation, and this property was utilized by those investigators to probe the structure of mitochondrial F_1-ATPase, indicating that this analog will interact properly with at least some enzymes. Under appropriate circumstances, the protecting moiety may also serve as a label.

Very recently, B Canard and R.S. Sarfati[2] have described similar nucleotides, here comprising all four nucleobases, with chemically removable 3'-hydroxyl protecting groups. Said protecting groups comprise various fluorescent dye moieties. These investigators have shown that these compounds may be added to appropriately primed polynucleotides by polymerases according to Watson-Crick base-pairing rules, and serve to terminate chain elongation in a manner which may be reversed by removal of said protecting groups by appropriate chemical treatments, admitting resumption of polymerization. These workers propose that such compounds may form the basis of a novel sequencing methodology availing stepping control by means of said removable protecting groups and detection of labels following their release from the nascent strand by appropriate chemical treatment. Such a method, while a potential advance over electrophoretic resolution methods, does not avail of great parallelism because only one molecule or an identical population of molecules may be sequenced at once (within a single vessel) by such a method, due to the

release of the labeling moiety prior to detection, according to this proposed scheme. Further, this limitation requires that any attempt to avail of parallelism entail elaborate parallel fluidics. Low or no parallelism entails that stepping rate will be critical to the throughput attained with such a sequencing scheme. The results published by these authors suggests that the rate of chemical removal of 3'-hydroxy protecting groups (less than 90% removal after 10 minutes of treatment with 0.1M NaOH) will be unacceptably low for such an inherently serial sequencing scheme.

Additional references regarding such compounds and in most instances their properties as substrates for various enzymes including polymerases have been found in the biological literature:

Churchich, J.E.; 1995. *Eur. J. Biochem., 231*:736.

Metzket, M.L.; Gibbs, R.A.; et al.; 1994. *Nucleic Acids Research, 22*:4259.

Beabealashvilli, R.S.; Kukhanova, M.K.; et al.; 1986. *Biochimica et Biophysica Acta, 868*:136.

Chidgeavadze, Z.G.; Kukhanova, M.K.; et al.; 1986. *Biochimica et Biophysica Acta, 868*:145.

Hiratsuka, T; 1983. *Biochimica et Biophysica Acta, 742*:496.

Jeng, S.J.; Guillory, R.J.; 1975. *J. Supramolecular Structure, 3*:448.

Related Base Addition Sequencing Schemes:

Various other investigators have also independently devised polynucleotide sequencing methodologies which depend on the addition of a polymerization terminating labeled nucleotide to a primed or elongated daughter strand on a polynucleotide sample with template dependent polynucleotide polymerases. Most, but not all, of these methods (referred to herein as previously disclosed base-addition sequencing schemes) avail nucleotide triphosphate monomers with some base-specific label which may be removed by some deprotection treatment. It must be emphasized that all of these other previously disclosed base-addition sequencing schemes examine not single molecules individually but rather large homogeneous populations of substantially identical molecules, wherein the observed signal used to identify label type originates from the totality of such a population of molecules rather than an individual molecule. It must be further emphasized that conventional usage does not generally reveal this distinction: phrases such a "a molecule" or "a sample molecule" refer not to an individual molecule considered separately or in isolation from other molecules including separately from other molecules of identical composition and structure, but to populations comprising millions or more molecules of identical structure. A careful reading of these prior

disclosures reveals that these investigators are not working with samples consisting of single molecules but rather with samples comprising a plurality of identical molecules. In particular, even where these investigators do not (as is consistent with conventional usage) explicitly note this point, they take measures which would apply only to samples of pluralities of identical molecules, and do not take measures associated with working with single molecules.

Disclosures of this type, yielded in a search, are:

Patent Number:	Inventor:	Issued:
US 5,302,509	Cheeseman, Peter C.	12 April 1994
WO 93/2134	Rosenthal, André; et al.	28 October 1993
DE 41 41 178 A1	Ansorge, Wilhelm	16 June 1993
WO 93/01583	Gibbs, Richard A.; et al.	18 March 1993
WO 91/06678	Tsien, Roger Y.; et al.	16 May 1991
WO 90/13666	Garland, Peter B.; et al.	15 November 1990

Included in some of these above disclosures are descriptions of nucleotide triphosphates comprising removable fluorescent 3' protecting groups.

Object of the Invention

The present invention provides methods of detection and discrimination which address the complexity found in biological systems, though they may further be applied to non-natural systems including but not limited to mimetics. Much of this complexity derives from combinations or permutations of simple units such as the four nucleotide bases of polydeoxyribonucleic acids and polyribonucleic acids, or the twenty common amino acids found in polypeptides and proteins.

This complexity, which underlies the most diverse and nuanced of biological processes, has presented both the promise that ultimately much mechanistic knowledge of biological processes may be gained through the accumulation of greater information about underlying structures and biopolymer sequences, and the correspondingly motivated challenge of full enumeration and determination of these structures and sequences.

Because typical eukaryotic genomes contain between 10^7 and 10^{10} DNA base pairs, and because there are several well studied organisms of particular interest, economical and technically simple methods capable of determining the full genome sequence of an individual organism over a conveniently short period of time would be particularly desirable.

The availability of such sequencing methods would enable greater clinical applications of molecular medicine, would facilitate greater and safer application of gene therapy, would permit timely completion of the several genome projects within fiscal constraints, and would enable facile gathering of genome information on populations of individuals, which would have applications in such areas as the study of polygenic diseases, epidemiology and field ecology. Such applications are presently limited by the cost and cumbersome nature of existing sequencing methodologies.

Combinatorial chemistry, affinity characterization, therapeutic synthetic immunochemistry, pharmacology and drug development, *in vitro* evolution and other fields concerned with the elaboration of a diverse population of molecules, their characterization according to desired properties, and recovery or identification of molecules displaying suitable characteristics may be favorably improved by the availability of methods which permit the introduction of and both qualitative and quantitative characterization of kinetic and equilibrium properties of molecular recognition and binding phenomena, particularly where such parameters may be used as selective constraints.

There has further been some interest in rebuilding or supplementing the immune systems of immunocompromized individuals, and in the development of highly specific antibiotic agents targeted to antibiotic, antifungal or antiviral resistant or otherwise poorly treatable pathogens. Both of these goals may be furthered by the use of the methods of the present invention as they may readily be applied to the determination of pathogen specificity and antigenicity.

Summary and basic description of the Invention:

The present invention approaches the vastness of biological complexity through massive parallelism, which may conveniently be attained through various single molecule examination (SME) methods variously referred to heretofore as single molecule detection (SMD)[3], single molecule visualization (SMV) and single molecule spectroscopy (SMS)[4,5] techniques. Used within appropriate procedures, single molecule examination methods can enable molecular parallelism.

Molecular parallelism may be applied to the examination of the composition of complex molecules (including co-polymers of natural or of synthetic origin) or to determinations of interactions between large numbers of molecules. The former case may be applied to genome-scale sequencing methods. The latter case may be applied to rapid determination of molecular complementarity, with applications in (biological or non-biological) affinity characterization, immulogical study, clinical

pathology, molecular evolution (e.g. *in vitro* evolution), and the construction of a cybernetic immune system as well as prostheses based thereupon. In both cases, molecular recognition phenomena are observed with molecular parallelism.

Note that within said affinity characterization applications, both kinetics of both binding association and dissociation, and binding equilibria, may be examined. Kinetics may be examined by observing the rates of occupation of appropriate sites or diverse populations thereof by some homogenous or heterogeneous sample, and the rates of vacancy formation from occupied sites. Equilibria constants may be determined by observing the proportion (number of occupied sites divided by number of total sites) of sites occupied under equilibrium conditions, with greater quantitative confidence yielded by, for example, examining more binding sites.

Sequencing of polynucleotide molecules may be effected by the (preferably end-wise) immobilization of a library of such molecules to a surface at a density convenient for detection, which will vary according to the detection methodology availed. Several methods capable of effecting such immobilization will be obvious to those skilled in the arts of recombinant DNA technology and molecular biology, among others. Priming, which may be random or non-random, is effected by any of a variety of methods, most of which are obvious to those skilled in the relevant arts.

Genome sequencing applications availing of enzymatic polymerization's, and corresponding embodiments of the present invention, rely upon control over polymerization rate and nucleotide incorporation specificity, consistent with the well-known Watson-Crick base pairing rules which may be enforced (upon single nucleotides in a processive manner, as conditions permit) by the use of DNA polymerases or analogs thereof, in combination with repeatable single molecule detection applied to a large population of diverse molecules. A sequencing cycle comprises the steps of: (1.) polymerizing one or less nucleotides, which carry some removable or neutralizable molecular label and may optionally be reversibly 3' protected (or otherwise protected in any manner which modulates polymerization rate) onto each sample molecule at the primer or at subsequent extensions thereof, and in opposition to (and pairing with) a single, unique, base of the template polynucleotide strand; (2.) optionally washing away any unreacted labeled nucleotides; (3.) detecting, by either direct or indirect methods, said labeled nucleotides incorporated into said sample molecules, in a manner which repeatably associates information obtained about the type of label observed with the unique identity of the template molecule under observation, which may be uniquely distinguished by a variety of methods (which include: a mappable location of immobilization of the sample

template molecule on a substrate surface; a mappable location of immobilization of the sample template molecule within some matrix volume element; microscopic labeling with some readily identifiable, e.g. combinatorially or permutationally diverse and readily examined particle or molecule or group of molecules and detection of the thus marked identity of individual free molecules in solution; and, scanning of a liquid sample volume where sample molecules and sample conditions are matched to ensure manageably slow free diffusion of sample molecules permitting tracking of the motions of free individual molecules in solution, in which instance unreacted labeled monomers may be removed, for instance, by filtration), and recording the information obtained by said detections, especially in a manner which may be conveniently recalled according to the unique identity template molecule to which it refers; (4.) removing or neutralizing said incorporated label; (5.) optionally removing (by appropriate means) any 3' protecting groups (or any other protecting groups which may serve to modulate monomer addition rate to the strand being copied from the template molecule) from the nucleotide added during the present cycle, if these are distinct from any cleavably linked labeling moieties; (6.) optionally checking that the removal or neutralization of said label in step (4) was successful for any particular molecule of the sample, by repeating a similar detection procedure. Said sequencing cycle comprising an appropriate subset of steps 1-6 may be repeated as many times as convenient, but must be repeated a sufficient number of times to obtain sequence information of sufficient complexity from each individual molecule to permit unambiguous alignment of all such sequence information determined for all of the molecules of the sample. This minimum number of cycles may be approximately related to the complexity C of the sample to be treated as part of the same macroscopic reaction (i.e. a macroscopic sample preparation subjected to unitary macroscopic manipulations) by the formula $C<4^n$ where n is the number of cycles. Beyond this minimum, there are tradeoffs between the number of cycles to be performed and the number of molecules to be examined, and the confidence for sequence data obtained. Note that unused reagents and enzymes may be recovered from washes and recycled.

In contrast to the previously disclosed base-addition sequencing schemes, the sequence determination applications of the present invention enjoys substantial advantages deriving from sample manipulation in the single-molecule-regime. Working instead in the distinct single-molecule regime rather than with populations of identical molecules provides substantial advantages of parallelism, facility of use and implementation (including automated implementation,) and operability. Among these are

unanticipated advantages: (1) because a single molecule is necessarily monodisperse, failure of a molecule to undergo addition in a cycle does not cause a loss of sample monodispersion (i.e. lead to uneven sample molecules dispersity or polydispersion); such addition failure is unproblematic when single molecules are examined individually because it is readily detected and accounted for in data analysis; in contrast, samples comprising multiple identical molecules may thus take on non-identical lengths, complicating data collection and analysis; (2) samples comprising a plurality of individually distinct single molecules (species) may be handled unitarily without requiring any handling measures to keep distinct molecules apart, providing a large reduction in manipulations required on a per-species basis and not requiring the use of many separate, parallel fluid handling steps or means; (3) inadvertent multiple base additions are more readily detected and their extent is more readily quantified because these changes in quantity are large compared to the signal expected from the incorporation of a single base (i.e. single label) into a single molecular species; (4) deprotection or delabeling failures may also be readily detected and noted for the correct single molecule, such that addition failure, the presence of a label, or overlabeling in the subsequent cycle may be correctly interpreted (according to the unlabeling and single stepping methods used in a particular embodiment.) These advantages are expected to be important in the competitiveness of these present methods over conventional polynucleotide sequencing methods.

Various techniques are included to address any non-idealities encountered which may arise because of deviations from conventional polymerization or detection methods. These generally take the form of different types of redundancy, which may be employed to either prevent or resolve any such errors. Prominent among these redundancies is oversampling, i.e. the examination of some multiple (j) of the number (m) of sample molecules suggested by combinatoric computations to be minimally sufficient for full alignment of data from a sample of a given complexity. Such oversampling redundancy will increase the confidence interval for accuracy of collected data and reduce the likelihood of artifacts arising from sequence duplications which may occur in any given sample. Oversampling redundancy may be availed to increase data confidence by providing the opportunity to score and match multiple occurrences of the same sequence segment and thus detect and eliminate erroneous sequence segment information by virtue of its less frequent occurrence. Erroneous sequence segment information may arise, for instance, by nucleotide incorporation errors which are an inevitable feature of polymerization with polymerases having a characteristic fidelity, i.e. displaying a

characteristic nucleotide misincorporation rate. Such methods will be particularly useful where polynucleotide polymerases fidelity would otherwise be unacceptably low. It should be noted that an error rate of one percent or more has been deemed conventionally acceptable for genome informatic purposes.

Further, known molecules having sequences that are highly unrelated to the sample may be included as internal controls to monitor the efficiency and accuracy of a particular sequence collection process; such internal control sequences will present negligibly small overhead because molecular parallelism may easily accommodate any such comparatively small increase in sample complexity, even though it might be considered large with respect to pre-existing methods.

After raw data have been collected for each molecule, these are all mutually compared by some appropriate matching algorithm and aligned so as to reconstruct the full sequence of the sample. The computational complexity of completing such an alignment may be estimated as the multi-phase comparison and sorting of (j)(m) strings each of length n. Alternatively, data alignment may be performed in tandem or parallel with later cycles and may be monitored by appropriate computational algorithms for data quality and confidence of sequence information, and cycling may continue till desired criteria are satisfied. Computer, microprocessor, electronic or other automated control of instrumentation, including fluidics and robotics for the manipulation of samples, and the automated effectuation of the various methods of the present invention, all according to parameterized algorithms, may be accomplished by means obvious from the present disclosure to those skilled in the relevant arts (e.g. fluidics, robotics, electronics, microelectronics, computer science and engineering, and mechanical engineering). Concurrent data alignment and monitoring will permit modifications of the sequencing cycle described above, such as dynamic adjustment of polymerization reaction conditions and durations, label removal or neutralization procedure parameters, polymerization deprotection conditions, and any other desired parameter, so as to permit optimization of procedures and results.

With appropriately flexible design, automated systems and instruments such as those described above for genome applications may readily be adapted, with appropriate changes in samples and labeling methods and reagents, to cybernetic molecular evolution, cybernetic immune system, broad spectrum pathogen characterization and other applications of the present invention.

According to the embodiment availed, double or single stranded polynucleotides may be examined. Where single stranded polynucleotide molecules are preferred, second strands may be removed by performing said immobilization so as to only involve only one strand in covalent linkage with said surface and then performing a denaturation of the sample with washing. Priming means required by any particular enzyme must then be provided, usually by hybridization of a complementary oligo- or polynucleotide to the sample template molecules, though other means are possible. Other methods which will be obvious to those skilled in the arts of recombinant DNA technology may also be employed to yield immobilized or otherwise uniquely identifiable single stranded polynucleotide samples. Where double stranded molecules are preferred, said second strands may be treated with an appropriate exonuclease under appropriate conditions and for an appropriate lengths of time to provide a good distribution of lengths of said second strands such that the termini of the undegraded portions of said second strands provide convenient priming for enzymatic nucleotide polymerization (i.e. DNA directed DNA synthesis or DNA replication, DNA directed RNA synthesis or transcription, RNA directed DNA synthesis or reverse transcription, or RNA directed RNA synthesis or RNA replication).

Note that the polynucleotide sequencing methods of the present invention represent the converse of conventional enzymatic and chemical sequencing methods in that those conventional methods rely upon the production of multiple homogeneous sub-populations of DNA molecules which together comprise a nested set, and the detection of each of such sub-population (with deviant chain terminator misincorporation molecules arising with significantly lower frequency and thus constituting a poorly detected population), while the present invention relies on alignment of information from a highly inhomogeneous population molecules and repeatable detection of single molecules. Further note that by previous methods, each species yields information about only one base at one position within the sample sequence, while with the methods of the present invention, each individual sample template molecule may yield information about the identity of several bases. Note also that under conventional methods, some effort has been expended in increasing the number of bases yielding information per sample, i.e. lengthening the linear sequence information obtained from any one segment of a sample, which is substantially frustrated by the inherent limitations of electrophoretic separation and particularly gel electrophoresis, while the present invention readily accomplishes the

information yielded per unitary manipulation through increases in the facility and practicable extent of parallelism.

There are several levels of parallelism and pipelining possible with the methods of the present invention. An arbitrarily large number of molecules may be subjected to any given manipulation at once if they are part of the same unitary sample. Detection will have constraints entailed by the particular instrumentation and method used, but many degrees of freedom exist with regard to means of providing parallelism in detection instrumentation (e.g. multiple microscopy instruments or appropriately arranged objective lenses and controlled light paths for light microscopic based detection, multiple optoelectronic device arrays [e.g. CCDs or SLMs] for the respective types of detection; multiple probes [i.e. in arrays with parallel detection provided] for scanning probe microscopic detection methods with various degrees of freedom with respect to eachother during scanning, etc.) Means for pipelining the steps of the methods disclosed herein will be readily apparent when one considers that dedicated instrumentation or robotics may perform each relevant step, and that the ensemble of such instrumentation may readily be integrated to form a coordinated system, for example matching throughput at different stages by adjusting the parallelism of appropriate stages. Thus economy, throughput and data accuracy are tradeoffs, but may individually vastly exceed any such measures attainable with conventional methods.

Detailed Description of the Invention:

Definitions:

For purposes of the present disclosure and invention the term nucleotide shall comprehend both the standard four deoxyribonucleotides or the standard four ribonucleotides, as well as any variations thereupon which pair with similar or other bases according to definable pairing rules. The phrase nucleotide analog shall refer to any naturally occurring or synthetic variation thereupon which may be utilized within the unity of the present invention to gain sequence information about a given polynucleotide sample.

Unless otherwise indicated, nucleotides are all 5' triphosphates. Phosphorothioate derivatives of nucleotides may be employed interchangeably with these without departing from the essence of the present invention. Said phosphorothioates provide the additional advantage upon polymerization of yielding polynucleotide backbones which are less susceptible to the intrinsic exonuclease activity of various polynucleotide polymerases. Alternatively, polymerases with no detectable intrinsic exonuclease

activity, or with no detectable 5' to 3' exonuclease activity, may be favorably used.

It is further understood that unusual or synthetic polymerization chemistries will also be comprehended within the unity of the present invention, provided template directed pairing is preserved.

While well studied DNA polymerase enzymes, preferably lacking a 3' to 5' exonuclease activity, or RNA polymerases or reverse transcriptases may be initially preferred for use in sequencing applications of the present invention, use of the term polymerase (as well as the term transcriptase) shall refer to any molecule or complex capable of enforcing fidelity of pairing on single nucleotides at a structurally defined site of a template polynucleotide molecule, whether or not said any molecule or complex itself catalyses the addition of said single nucleotide to said structurally defined site or not, and whether said any molecule or complex is a naturally occurring enzyme or ribozyme or sythetically derived catalyst such as an abzyme with the appropriate catalytic specificity or an artificial receptor molecule or any purely synthetic assemblage capable of enforcing high fidelity of comonomer association (e.g. base pairing) according to well defined rules.

Methods for repeatable detection and identification of single molecules:

Repeatable detection and identification of single molecules is achievable by microscopic labeling with some readily identifiable, e.g. combinatorially or permutationally diverse and readily examined particle or molecule or group of molecules and detection of the thus marked identity of individual free molecules in solution, with removal of excess nucleotides (e.g. by filtration); and, scanning of a liquid sample volume where sample molecules and sample conditions are matched to ensure manageably slow free diffusion of sample molecules permitting tracking of the motions of free individual molecules in solution, as observed by T.T. Perkins et al. for reptation of DNA in solution[6,7], in which instance unreacted labeled monomers may be removed, for instance, according to their more rapid diffusion, possibly through a filter, and detection may favorably comprise observation of reduced mobility of a labeling moiety after it has become attached to a sample molecule.)

According to the labeling methods employed, various detection methods may satisfy the requirements of signal detection with repeatable assignability to a particular unique sample template molecule.

Prominent among these detection methods are microscopy methods such as: video microscopy including confocal fluorescence microscopy with or without

enhancement, and with or without variations incorporated into the present invention: near field scanning optical microscopy (NFSOM)[8] and variations thereof; contact and non-contact varieties of scanning force microscopy (SFM; also termed atomic force microscopy [AFM])[9] and variations thereof; other scanning probe microscopies including scanning tunneling microscopy (STM), scanning tunneling spectroscopy (STS)[10], and so-called field emission mode STM (which is more accurately described as microscopy by field emission from a scanned conductive probe, or scanning field emission microscopy, SFEM, because no tunneling actually occurs)[11]. Any enhancements of scanning probe microscopy, including multiple probe parallelism, may readily be availed in the practice of the present invention.

Additionally, optical detection methods employing optoelectronic array devices (OADs), such as spatial light modulators (SLMs), laser diode arrays (LDAs), light-emitting diode arrays, or charge coupled photo-diode arrays (conventionally termed CCDs), in combination with appropriately high sensitivity detection methods, may also be employed, particularly with samples immobilized such that the maximal proportion of pixel elements of said array will be involved with the detection of a signal from exactly one sample molecule. CCD and SLM array device are presently available at pixel densities of approximately 10^5 to 10^6 per cm^2. LDAs of comparable density are currently under development. Device level constraints upon parallelism will thus be significant , but may be overcome by increasing the data obtained per molecule (i.e. processivity or sequence segment length.) Such devices may be employed remotely, i.e. in some arrangement where light passes through the sample under study and is detected by some apparatus involving said array devices, or in close or direct contact with said sample, as for instance, polynucleotides have been immobilized to integrated circuits for other applications. Appropriate arrangements of such devices for the appropriate detection scheme in which each device type is appropriately used will be obvious to those skilled in the arts of optics and optoelectronics.

Note that for purposes of those variations of the present invention involving the immobilization of sample molecules, said immobilization may be conveniently effected in a random manner, relying upon some appropriate surface or volume density which yields a corresponding random surface or volume distribution, and appropriate detection methods to permit repeatable resolution of most sample molecules from eachother. The length of the molecules in question will be an important factor in the determination of a desirable said density. Generally speaking, for random surface immobilization and without the use of measures to orient or order sample

molecules, for molecules of length L (which may additionally account for any labeling bead diameter),and detection methods relying on spatial resolution R, maximum practical molecule number density will generally be the less than $1/(2L+R)^2$. This assumes the worst case configuration of two end immobilized molecules extending directly towards eachother and both labeled near their respective termini. Similar calculations may be applied to three dimensional cases. Alternatively, one may consider $(2L+R)^2$ or $(2L+R)^3$ to be an average bin size, and determine via the Poison distribution the optimal molecular number density corresponding to the largest number of bins being occupied by precisely one sample template molecule.

Alternatively, molecules may be labeled by a first label, for example with a particular fluorescent dye incorporated by nick translation, in a manner identifying a portion of the molecule near the site of polymerization, and proximity of said first label to the perceptibly distinct labeling moieties used for nucleotide incorporation detection and discrimination will permit the detection of unacceptable proximity of two distinct sample molecules. Such a method is consistent with the tracking methods described below for free sample molecules. In such a case, the data collected during the cycle in which said unacceptable proximity is observed for the affected molecules may be ignored, and lack of information from this cycle noted for the respective molecules. Conditions, such as solution viscosity, sample molecule diffusion rate, sample molecule concentration, sample dimensions, etc., may be optimized to reduce the occurrence of such unacceptable proximity, and oversampling methods described in other portions of the present disclosure may be applied to preclude this form of error from degrading final data quality. These methods may be applied to either immobilized or unimmobilized sample molecules.

Microscopy based detection:

Light microscopic visualization represents a particularly convenient and technically simple detection and unique molecule localization method. A visualization method of particular interest for purposes of the present invention in higher performance or more demanding applications is video enhanced confocal fluorescence microscopy (VECFM)[12], preferably utilizing optics well matched to the refractive index of the reaction or detection medium.

As discussed above, various scanning probe microscopies may also be advantageously used within the present invention according to labeling agents and methods used. Most prominent among these are NFSOM and

variations thereof, and both contact and non-contact SFM, and variations thereof.

Generally speaking, a microscopy based detection method must be sufficiently convenient, capable of use with a stepper translated or otherwise translatable sample, not destructive of the sample, and capable of detection of any labeling methodology to be used with it. Thus, it is quite likely that many microscopy methodologies not yet developed may readily be employed with the present invention. Further, microscopy and corresponding apparata shall comprehend any miniaturized or microfabricated microscopy devices or other comparable integrated detection means.

High sensitivity and Scanned Excitation Beam Fluorescence confocal microscopy:

A modification of VECFM which is particularly suited for SMD and SMV relies upon selective fluorescent excitation of an appropriate dye molecule label (or of molecules within a sample with appropriate fluorescent properties independent of labeling) in some sample by means of some tightly defined beam, with dimensions at or near the resolution limit of the apparatus, of an appropriate frequency, or of parametrically controllable frequency, where said beam is caused to scan in a controlled manner through the sample region within the visual field. This microscopy, including numerous variations, may be termed either scanned beam confocal microscopy or steered beam confocal microscopy (in either case, SBCM). Scanning of said beam through the sample within the visual field may be accomplished by introducing said beam into the optical path of the VECFM via mobile mirrors which may effect said controlled scanning, or by first producing said beam with a pinhole which is itself scanned, before deflection towards the sample via said mirrors, which in the present case may be fixed in position, through the use of pinholes in a rotating disk arranged in one or more spiral arms to effect an approximately rastering illumination of the sample as said disk rotates, or by other means which will be obvious to those skilled in the design of optical instrumentation and microscopy. Said beam will excite fluorescence in any appropriately responsive molecules which occur in its path. An optical splitter may then redirect a fraction of the light transmitted from the sample through the objective lens, and direct it through a narrow bandwidth, high transmissiveness filter, which may be specific for a fixed or for a parametrically controllable variable frequency, to uniquely select the appropriate fluorescent emission frequency, to a highly sensitive photodetector, which may record either intensity as intensity information or as the number of photons detected per unit time, as a function of the region being subjected to fluorescence exiting illumination or being distinctly observed (see

below). Thus a high resolution map of the fluorescence of the sample may be reconstructed, and further overlayed images obtained for the same sample and sample location by conventional VECFM means.

Alternatively, the entire sample of visual field may be subjected to illumination by an appropriate excitation frequency, and a pinhole scanned through the portion of the output of said optical splitter, such that light passing through said pinhole will reach said highly sensitive photodetector.

In yet a third, albeit technically more complex implementation, an SLM, may be used in place of said pinhole (in either configuration), and fluorescent excitatory illumination may be either broadly distributed or scanned.

In a fourth, albeit technically more complex implementation, sensitive photodetection may be accomplished with a highly sensitive CCD, and fluorescent excitatory illumination may be either broadly distributed or scanned. At present, CCD sensitivity approaching single photon detection is technically possible though is not practical for high volume applications.

In a fifth implementation, said scanned beam may originate from a laser diode array device or a light emitting diode array device, where only one of, or a contiguous group of elements of, such an array is active at any particular time so as to produce a particular beam, and the group of active elements of said such an array is changed as a function of time to effect scanning of the sample by the coordinated activation and deactivation of the plural beams thus produced.

In all of the above implementations, spatial information is gained about any particular fluorescent emission, and this may then be combined with other visual information obtained via the same VECFM apparatus.

Note that for scanned beam methodologies, where beams are used for excitation or detection, even where said beams may have inhomogeneous but invariant distribution of internal flux density, known samples such as individual dye molecules may be imaged for calibration purposes and information useful for algorithmic enhancement may be collected. This information represents the characterization of the convolution of the beam and optics properties with the signal actually owing to the known sample, and thus localization of fluorescent sample features may be accomplished at better than optical resolution limitations. For example, a single, immobilized fluorescent molecule may be examined by such an apparatus, and the intensity as a function of beam position may be recorded for the full duration of its presence within the beam's path as said beam scans the sample, and the data thus obtained may then be used to determine the change

in observed intensity as the sample molecule enters the extremity of the beam, traverses the beam and exits the beam. This information may then be subjected, for instance to averaging or other computations to determine the relationship between the location of the molecule within the beam and the intensity observed, and finally that information used to estimate the intensity which would be observed when such a calibration sample molecule is in the precise center of the beam. This information may then be used in image enhancement of unknown samples. Note, however, that localization to below optical resolution limitations is distinct from increasing the resolution capability for two nearby objects.

Scanning beam microscopies will be of particular advantage where it is desirable to use particular illumination frequencies to modify the sample. For purposes of the present invention, a beam of predetermined frequency, for instance delimited and scanned by means of a pinhole as described above, may be used to selectively modify a particular sample molecule. For example, a beam of predetermined frequency may be used to effect the photobleaching of the labeling moiety on a particular sample molecule. to selectively remove a photocleavable protecting group on a particular sample molecule, to selectively remove a moiety joined to a sample molecule by a photocleavable linker, or selectively control any photochemical reactions in a highly localized but non-invasive manner.

Note that implementations permitting variations of illumination frequency and/or variations of the frequency or frequencies selected by filters for detection purposes constitute microspectroscopy or microfluorimetry, and may be applied to any of the various light microscopies.

Repeatability by immobilization with discernible location:

Surface Immobilization:

A large number of methods presently exist to effect the immobilization of macromolecules and other molecules to various surfaces including the surfaces of optically transparent materials. In general, such methods rely on the chemical modification of said surfaces such that they will be reactive with or have specific affinity for particular chemical functional groups placed on said macromolecules or molecules.

Applicable methods include those described by S.P.A. Fodor et al.[14] to effect micropatterned surface immobilization and controlled synthesis of polypeptides and polynucleotides, those described by M. Hegner et al.[15] to effect the end-wise immobilization of terminally thiol modified double helical DNA molecules to a gold coated surface, or those methods recently used by L. Finzi and J. Gelles[15] to effect end-wise attachment of DNA

molecules to an antibody coated glass surface. Many alternative methods will be obvious to those skilled in the relevant arts.

For purposes of genome sequencing applications of the present invention, DNA from a cosmid library which may have been prepared from total genomic material, from a cDNA library derived from a particular tissue type, from a cosmid library which may have been prepared for a single chromosome or group of chromosomes or particular chromosome segments, or directly purified genomic DNA or directly purified RNA from a particular cell type, etc., may be subjected to fragmentation. Physical methods such as shearing with a hypodermic apparatus may be suitable. Where the sample is in the form of duplex DNA, it may be treated with restriction enzymes, which preferably restrict either 6- or 4-base recognition sequences, so as to produce sample molecules of mean length of either 4 kilobases or 256 bases, respectively. Such lengths are sufficiently short to yield a high number density of sample molecules. Said sample molecules may then be appropriately derrivatized, for example by fill-in reactions at 5' overhang cohesive termini produced by said restriction enzymes with nucleotides bearing an affinity label or an appropriately reactive chemical functional group.

Matrix Immobilization:

There has been increasing interest and progress in the field of affinity chromatography which relies upon varyingly specific affinity interactions between molecules immobilized to a chromatographic matrix or polymeric matrix and the molecules contained in some sample. Of particular relevance are matrices with polynucleotides immobilized thereupon. An example which is widely known and used within the relevant fields is oligo-dT cellulose. Further, many chemistries and methods used to immobilize macromolecules to surfaces will be similarly applicable to immobilization to a polymeric matrix provided said matrix is chosen so as to have appropriate reactivities and not pose any difficulties associated with non-specific interactions. Most methods capable of effecting such matrix immobilization will be acceptable for purposes of the present invention. Note, however, that any matrix used in the present invention must admit the sufficiently rapid transport or diffusion of reagents, enzymes and buffers, as required by the particular embodiment.

Focal plane scanning:

For detection an discrimination within a volume, whether for matrix immobilized samples or diffusion constrained free molecules in solution, especially where fluorescent labeling of one form or another has been

employed, a sample may be examined by microscopy with reconstruction of three-dimensional spatial information by scanning the focal plane through the depth of the sample and collecting image data at appropriate intervals. Such methods of three-dimensional reconstruction are well known within the art of microscopy.

Plane Excitatory Illumination:

Alternatively, optical means such as moving slits or SLMs or laser diode arrays may be employed to selectively illuminate a particular region, preferably a single plane (of thickness similar to the wavelength of light employed or feature size of integrated device means employed), to examine a particular subset of sample template molecules and labels associated with them, providing spatial reconstructability of the data thus collected.

Two Beam methods including plane illumination:

Volume distributed samples may also be examined with methods closely analogous to those recommended for three dimensional optical mass data storage, for instance, by Sadik Esener in U.S. Patent Number 5,325,324. Here, labels requiring excitation by photons of two distinct frequencies for photoemission may be employed. Alternatively, the related methods of illuminating an entire plane of a sample with one of said distinct frequencies may be availed as a mechanism for imaging with spatial reconstructability.

Immobilization via concatenation:

For the various applications of the present invention involving the interaction of enzymes with extended linear macromolecules such as polynucleotides, when said extended linear molecules may be conveniently circularized by appropriate treatments (which will generally be obvious to those skilled in the relevant arts), immobilization of said extended linear molecules may be conveniently effected by their concatenation with second extended linear molecules which are likewise conveniently circularized by appropriate treatments (which will again generally be obvious to those skilled in the relevant arts) bearing chemical properties (i.e. functional groups such as thiols or affinity moieties such as biotin) favorable for convenient, specific immobilization to a surface, matrix or other solid support. For purposes of, for example, certain sequencing applications of the present invention, said second extended linear molecules are favorably bound (with methods which will generally be obvious to those skilled in the relevant arts) at a predetermined location along their length, to some protein, which may be an enzyme such as a polymerase, before immobilization. Said second extended linear molecules may have termini

with reactive chemical functional groups which may be bound together by the addition of some appropriate reagent such as a chemical cross-linking agent, or with some affinity moiety such as an oligo- or polynucleotide which may be bound together by an appropriately complementary oligonucleotide or polynucleotide (with or without ligation thereof), or some appropriate multifuctional binding protein or receptor. Such an arrangement permits the following steps to be performed: said second extended linear molecule is bound to said enzyme; said protein is caused to bind to said first extended linear molecule (which may be circularized either in a prior or subsequent step); said second extended linear molecule to which said protein has been bound is caused to circularize by appropriate treatment; and if said first extended linear molecule is at this stage linear, it is caused to circularize. Without any special measures, there is a fifty percent chance that such a process will result in concatenation of the first extended linear molecule with the second extended linear molecule. Numerous methods, such as size separation followed by retention by immobilization, may be used to purify the resulting desired concatenate. Where said second extended linear molecule was chosen to be relatively short, such an assemblage will provide for the retention of said first extended linear molecule, now in concatenated circular form, in proximity to said protein, with specific immobilization or convenient immobilizability. Thus, said protein and said first extended linear molecule now in concatenated circular form have a high effective concentration with respect to eachother upon dissociation, and said protein and said first extended linear molecule now in concatenated circular form will not interact with the molecules of other such assemblages when said assemblages are at sufficiently low density or said second extended linear molecule now in concatenated circular form is particularly short (i.e. effectively shackles said first extended linear molecule now in concatenated circular form to said protein whether or not said first extended linear molecule now in concatenated circular form is bound by said protein.)

Such an immobilization scheme will be particularly desirable in, for example, sequencing applications of the present invention where a polymerase must perform a cycle, in which it binds, modifies and releases a sample molecule, at a high rate. A particular instance in which such desirability obtains is for samples to be analyzed with long sequence segments (e.g. hundreds or thousands of bases) where dissociation of the polymerase is necessary to permit either 3' hydroxy deprotection (e.g. removal of a photolabile protecting group) and or labeling moiety removal by appropriate means. Note that by immobilizing the enzyme, and hence the

spatial location at which the labeling moiety first comes into physical communication with a sample molecule, the above stated limitation on sample molecule density may be overcome, with the new limit being that imposed by the detection method, thus increasing sample density and in some embodiments the parallelism that thence may readily be achieved with detection methods such as microscopy. It is therefore feasible, with such assemblages, to collect sequence data dynamically from each molecule at a rate approaching the limits imposed by the slower of: the characteristic nucleotide incorporation rate of the polymerase; or, the diffusion rate limit of nucleotide association with the nucleotide binding site of the polymerase (divided by four) when nucleotides are at a sufficiently low concentration that their presence as labeled but free molecules in the detection field does not interfere with the detection (which may be time averaged according to the particular instrumentation used) of incorporated labeled nucleotides, which concentration will be dependent in part on the geometry of the liquid volume; or, the maximum rate of single label detection (but note that such a rate need not be low because detection rate will increase for multimeric labels, which may be employed). Such an immobilization method will favorably be employed for embodiments locating sample molecules on or near the surface of a CCD or SLM. Note that kinetic control of polymerization rate (and hence stepping rate, e.g. by adjusting nucleotide concentration) is also enhanced by the use of such a concatenation methodology.

Immobilization with non-random distribution:

While the above methods are convenient precisely because they require only the simple optimization of sample molecule density, the resulting random distribution will less than fully utilize available substrate or matrix space and fewer than all sample molecules will be sufficiently well separated for unambiguous resolution of two adjacent sample molecules. Due to the inherent advantages provided by molecular parallelism, this will not in general be a significant constraint. For applications in which a high degree of instrumentation miniaturization is desired, however, a better effective density of usable sample molecules, distributed in either two or three dimensions, may be effected as needed by non-random immobilization methods.

One such random immobilization method may avail of the invention of N.C. Seeman, described in U.S. Patent Number 5,278,051, which provides a process for the construction of complex geometrical objects. These methods may be applied to the production of regular two- and three-dimensional molecular lattices from polynucleotide compositions. The process of this invention

may be extended by the incorporation of appropriate affinity groups at predetermined locations within the objects, which for present purposes may favorably be small ligands such as biotin or digoxigenin, which may then be used as the target for a sample molecule which has been terminally labeled by a similar small ligand which has subsequently been bound by (an excess of) an appropriate multimeric receptor. Said multimeric receptor will then recognize and bind the complementary small molecule ligand incorporated into the structure of said lattice, and thus effect sample molecule immobilization according to the non-random pattern predetermined by the precise structure of said lattice and the precise distribution of ligands thereupon. Note that because the objects provided by the invention of N.C. Seeman comprise polynucleotide structures, care must be taken in using such a sample substrate with the methods of the present invention to ensure that said objects will be stable to all treatments which are to be applied to sample molecules, including denaturation, exonucleolytic degradation, primer hybridization, exposure to active polymerases, etc.,. Generally, these constraints may be met by effecting topological closure of all strands such that no free polynucleotide terminus is carried on such a lattice, and no denaturation procedures will result in matrix dissociation; the methods of the invention of N.C. Seeman may be availed in a manner meeting these constrains.

Note that to ensure complete regularity of lattices constructed by such means, or any other molecular lattices which do not have complete internal rigidity, the extremities of these lattices may be bound to solid supports which are then positioned so as to apply tensile stresses to said molecular lattices which will enforce constraints limiting flexural internal degrees of freedom and enforcing substantial spatial regularity on sample molecule distributions.

Any other method which provides a regular array of binding sites to which sample molecules may selectively be associated will also suffice for the purpose of non-random immobilization of sample molecules in two- or three- dimensions for the present invention.

Note also that said appropriate affinity groups incorporated (directly or, by conjugation or other methods, indirectly) at appropriate sites in a lattice may be chosen so as to interact directly with polynucleotide sample molecules in a sequence dependent or independent manner. Sequence dependent affinity binding may be effected with oligonucleotides or analogs thereof capable of forming double-, triple- or quadruple helices with said sample polynucleotides, ribozymes, or sequence dependent binding proteins including but not limited to: transcriptional activators (e.g. TATA-Binding Protein), enhancers and repressors; integrases; restriction

enzymes; replicator proteins (e.g. DnaA); DNA repair proteins; anti-polynucleotide antibodies, RNA processing complexes (e.g. snRNPs); and RNA binding proteins all under conditions permitting desired selectivity, specificity or stringency but, where appropriate, preventing polynucleotide cleavage or degradation. Where sequence specific binding is desired, and hierarchially prepared lattices are used, the distribution of particular specificities may be controlled by the staged incorporation of said affinity groups at various hierachial levels of the synthetic procedure. This will permit classification of sequence data according to the location of the sample template molecule from which it is obtained in the lattice (i.e. on the surface or within the matrix). Sequence independent binding of polynucleotides may be effected by the use of proteins such as RecA, histones, U1, etc.,.

Repeatable identification of unimmobilized molecules:

Single molecule tracking with controlled diffusion:

For samples under continuous observation, e.g. continuously within as visual field of a video microscope, molecules may be perceptibly labeled, for example by perceptible microscopic beads or the incorporation of a first fluorescent label, and tracked by the use of image analysis algorithms. Said algorithms will recognize only the appropriate type of label and track the motions of the respective sample molecule as it slowly diffuses in solution, so as to permit the unambiguous direct correlation or assignment of the signal associated with the addition of a labeled nucleotide to said respective sample molecule. For these methods, nucleotide labeling does not necessitate the use of large beads or other complexes for detection. Instead, single or oligomeric fluorescent labeling moieties, or enzymatic label affinity conjugation are preferred, such that labels may be removed without greatly disturbing the trajectory of said respective sample molecules. Either the direct colocalization (to within the resolution of the imaging method) of nucleotide label with said first fluorescent label or reductions in the Brownian motion of said nucleotide label sufficiently near (e.g. closest to) said first fluorescent label may be exploited in the detection of nucleotide label incorporation.

Note that manipulation with a laser trap, as for instance described by T.T. Perkins et al. for reptation of DNA in solution.[16,17] may be employed with such free molecules.

Unique labeling of sample molecules and identification methods:

Various methods may be employed to uniquely label individual sample molecules. The complexity of such unique labels must be greater than the

number of sample molecules contained within a unitary sample preparation, such that any label is highly unlikely to occur more than once within said unitary sample preparation.

Labels may be visually discriminatable, or may be diverse affinity labels or combinations thereof. Labels of this type may conveniently be random combinations of some basis set of distinct labels, formed for example, by a random coupling or polymerization of such labeling moieties to a defined chemical site provided by chemical modification of sample molecules.

Visual labeling may be accomplished by the use of a sufficient number of distinguishable fluorescent dye molecules, or other visual labels, such that the presence or absence of association of any one of said distinguishable fluorescent dye molecules may comprise the state of a bit in a binary code. Such labeling is similar to the combinatorial encoding described by S. Brenner and R.A. Lerner[18], but differs in that: perceptible labels may be used for encoding; labels need not be genetic material or linear copolymers; where only unique identifiability is required, the label moiety employed for encoding may be synthesized separately and possibly randomly, and bound possibly randomly with sample molecules; the information contained by each labeling moiety need not depend on its precise spatial association with sample molecules, or its location within a sequence, only its sufficient proximity; and, because of such modes of independence between the encoding, which serves here only for purposes of unique labeling, difficulties which may arise for particular orthogonal polymerization chemistries of different copolymer types may be avoided either by separate synthesis. Alternatively, for biopolymers, and possibly for specifically encoded libraries, the use of specific enzymes which may for example ligate polynucleotides or polypeptides, may be used to specifically control reactions and prevent polymerizations of one biopolymer from affecting a second, linked biopolymer. Note that moieties different from biologically occurring comonomers may be used as encoding label moieties, via functionalization of appropriate biopolymer segments with such moieties, in synthetic manners which will be obvious to those skilled in the relevant arts, or may be used, similarly, as constituents of the random library thus encoded. This latter case is, for example accomplished with the use of multiple distinct short double stranded DNA molecules with appropriately complementary cohesive termini which each carry some particular affinity or photolabel type, and which may be ligated together in a manner stepped by the addition of appropriate adaptor linkers, even in the presence of other biopolymers (such synthetic methods being further favorably facilitated by the use of solid phase synthetic

methodologies). Depending on the sensitivity of the detection methods used, multimers of each single type of fluorescent dye moiety, or detectable multiplications of other photolabels, may be used to effect higher modulo coding of labels.

Encoding by synthesis with multimacromonomers:

Note that the labeling methods of the present invention suggest a convenient solution to the problem recognized by Brenner and Lerner[19], as limiting the facility of their encoding system, i.e. the requirement of separate distinct comonomer (or co-oligomer) type addition steps for each polymer type. This prevents the use of highly random (but step-controlled) synthetic preparation of such encoded libraries, because the information encoded is realized by individual preparative synthetic steps, i.e. all of the information content of the encoding is conferred upon these compounds by the intervention or agency of a chemist (or automated systems) at each step. Such encoded libraries, of either the sequence encoded or modulo encoded types, including compounds comprising more than two polymer types, may be prepared with the following stepped random method in one container (with or without the favorable use of solid phase synthetic methodologies). Note that the term random here refers to the mixture of two or more multimacromonomers in each addition step, such that addition to all compounds under preparation will occur in a random manner within the reaction mixture, in a manner weighted according to the relative concentration of each such multimacromonomer. Such multimacromonomers may also be used in more directly controlled addition schemes with advantages which will be obvious to those skilled in the relevant arts. Multimacromonomers comprising two or more monomer (or macromonomer) types (e.g. comprising an amino acid monomer and a trinucleotide oligomer, or an amino acid monomer, a trinucleotide oligomer and a fluorescent or affinity labeling moiety) may be prepared by joining some or all of said two or more monomer (or macromonomer) types by cleavable linkers such as those described in other sections of the present disclosure. Thus, each multimacromonomer may be added to compounds under synthesis by addition of one of the monomer or macromonomer types to the corresponding polymer or macropolymer types of said compounds under synthesis by appropriate polymer synthesis chemistry, followed by addition of some or all of each of the remaining monomer or macromonomer types to the respective corresponding polymer or macropolymer types of said compounds under synthesis by appropriate polymer synthesis chemistry. Control over the details of such additions may be effected by control over, for example, removal of distinct protecting groups from distinct polymer or macropolymer types of said

compounds under synthesis by appropriate polymer synthesis chemistry. Linkers or specific linker branches may be cleaved at appropriate steps or after synthesis has otherwise been completed. Thus, correspondence between the composition of each polymer or macropolymer type comprised within each molecule of the compound under synthesis (which final composition may vary widely from molecule to molecule of the compound under synthesis, but strictly observe the correspondence between composition of some or all of each of the polymers or macropolymers comprised within each molecule of the compound under synthesis) is provided by the communication of the distinct monomer or macromonomer types comprised within each multimacromonomer. The first bond formed between a first monomer or first macromonomer of a multimacromonomer and a molecule of the compound under synthesis will thus ensure that other monomer or macromonomer types of the multimacromonomer which will be added at the respective multimacromonomer addition stage will correspond to the identity of the first monomer or first macromonomer thus added. Thus correspondence of some or all of each of the polymer or macropolymer types of final compounds is enforced (by the communication effected by, for example, linkers) even where the composition of some or all of the polymer or macropolymer types is respectively random.

Preferably, such linkers (which may be multiply branched, each of such branches possibly comprising cleavable groups susceptible to distinct cleaving treatments) are held in communication with some or all of the two or more distinct monomer or macromonomer types (which are added to the compounds under synthesis with distinct and mutually non-interfering addition or polymerization, deprotection and/or activation chemistries, termed "orthogonal" chemistries in the respective art) by attachment to the protecting groups used to effect the stepping of additions of each such multimacromonomer.

Said diverse affinity labels may be used in conjunction with multiple affinity separation paths and nucleotide label detection that associates the detected said nucleotide label with the resolved location of the respective affinity labeled sample molecule, thus accomplishing the required assignment of detection and discrimination of the appropriate nucleotide label precisely to the correct respective sample molecule. Alternatively, said diverse affinity labels may be added to sample molecules so as to be independently recognizable by appropriate receptor molecules or other affinity means, each complementarity type of which is respectively labeled with some distinct independently perceptible label.

Such labeling methods permit the processing of samples in fluid flow based apparata without the loss of single molecule identifiability or assignability of results. Also note that manipulation with a laser trap,

as for instance described by T.T. Perkins et al.,[20,21] may be employed with such uniquely labeled molecules.

Note that a case of encoding of particular interest is that of a functional molecule coupled to an informational molecule which is sufficient to direct the synthesis of said functional molecule in an appropriate, (e.g. biological or biological derived) system. Libraries of polypeptides expressed on the surface of, for example, bacteriophages carrying genetic material specifying said polypeptides, have found great use in the *in vitro* selection of binding specificities.[22] Encoding which may additionally direct synthesis may be availed in the affinity characterization and molecular evolution applications of the present invention. The communication of a synthesis directing informational molecule (favorably DNA or RNA) with the correspondingly synthesized one or more functional molecules (generally a polypeptide) may be effected by the *in vivo* coupling or otherwise compartmentally enforced unique one-to-one corresponding coupling of said informational and said functional molecule. A particularly convenient instance of such a molecules comprises the fused expression of said functional molecule or molecules as segments of the terminal proteins of the informational molecules (i.e. DNA) of various virus (e.g. adenovirus) or bacteriophage (e.g. PRD1 or phi29) genomes. Alternatively, said functional molecules may be fused with some molecule which associates in a specific manner with said terminal proteins, and which has sufficient opportunity during its *in vivo* synthesis, without or preferably with concurrent viral or bacteriophage replication, to associate with the terminal protein of the genomic material which determines the composition of said functional molecules, such that upon purification or lysis functional molecules remain in communication with the genetic material that determines their composition. Because biosynthesis of functional and informational moieties may favorably occur within the confines of a single cell. cross-coupling of inappropriate molecules may be readily avoided. Alternatively, the communication between polypeptide and polynucleotide moieties may be effected with some intermediate snRNP or snRNP-like moiety, where such an intermediate moiety may be targeted on the one hand by an appropriate affinity characteristic of one or more polypeptides to which said functional molecules are fused, and on the other hand by a polynucleotide sequence complementary (according to appropriate rules for double-, triple- or quadruple- helix formation) with the polynucleotide moiety of said intermediate snRNP or snRNP-like moiety. Such complexes comprising an intermediate snRNP or snRNP-like moiety may also favorably be formed within the confines of a single cell.

Cybernetic molecular evolution and algorithm mediated cybernetic molecular evolution of phenogenocouples:

Such polynucleotide-polypeptide chimera, or other molecule types comprising thus communicating and informationally corresponding chimera (e.g. where the polypeptide moiety has further been subjected to post-translational modification such as specific glycosylation and has been associated by some method to the respective genetic material determining its composition, for example by the sorting of individual cells carrying said genetic material in the form of a DNA vector with terminal proteins and expressing and processing said polypeptide, into distinct wells or vessels followed by disruption of membranes such that terminal proteins fused with peptides having affinity for the particular polypeptide of interest may come into contact with the processed polypeptide of interest, comprising a method for the molecular evolution of multiple-biopolymer containing macromolecules), which may be termed phenogenocouples, may be used as sample molecules with the broad methods of the present invention to effect the affinity characterization (including either or both equilibrium and kinetic characterization of molecular recognition including catalytic recognition and catalysis) of functional moieties and then the characterization and transcription of informational moieties thus determined to be of interest. Where algorithms control such a process, cybernetic molecular evolution is embodied.

Selected informational molecules may be selectively replicated or transcribed by activatable (e.g. photodeprotectable and especially 3' hydroxyl photodeprotectable) primers with appropriate complementarity to some region which bounds the informational content specifying said functional molecule or molecules. Alternatively, immobilization of a sample to be subjected to such manipulations may be effected so as to comprise some photolabile linkage, which may then be subjected to selective photodegradation to effect specific release. For immobilized samples, informational molecules which carry the relevant genetic component of a phenogenocouple may thus be released by either of these methods either singly, or as the population of multiple such molecules simultaneously copied or otherwise released according to the pattern of deprotection.

Alternatively, successive generations of molecules need only be related informationally, by analysis of composition of one generation, by, for example, the massively parallel characterization methods of the present invention, followed by de novo synthesis of molecules carrying the desired complexity and diversity of the succeeding generation. This is a particular distinguishing feature of cybernetic molecular evolution;

selection, amplification and mutation may be directed strictly by algorithms which manipulate data gathered about one generation to determine the composition of a succeeding generation.

Released molecules may then be recovered for subsequent amplification, mutation and subsequent rounds of selection by similar or other methods, as will be obvious to those skilled in the art of *in vitro* molecular evolution.

Note that post transcriptionally modified polypeptide moieties or other phenogenocouples may also be selected and otherwise subjected to *in vitro* evolution by conventional means as well as by the massively parallel examination and modification methods of the present invention.

Because of the correspondence between the diversity generation and selection aspects of molecular evolution, and immunological recognition and memory, all of these methods may be directly applied to cybernetic immune system applications of the present invention.

Labeled reagents and signal amplification and elimination techniques:

The categories enumerated below are included for description and not limitation; other appropriate labeling methods will be obvious to those skilled in the arts of biotechnology, cell biology and cytology, microscopy, organic chemistry, biochemistry or recombinant DNA techniques.

Each category will comprehend a variety of specific variations, as will be obvious to those skilled in the relevant arts. Various labeling methods will generally correspond best to various detection methods.

Fluorescent labels:

Detection methods for the present invention may favorably exploit fluorescent labeling techniques.

Genome sequencing applications of the present invention may thus avail of established fluorescent modification and detection methods. Other applications of the present invention may also benefit from the application of fluorescence modification and detection methods.

Much effort has already been invested in the development of fluorescently labeled nucleotide triphosphate compounds and analogs thereof. Many such compounds are acceptable substrates for polynucleotide polymerase molecules. These compounds have therefore proven suitable for use in various electrophoresis based DNA sequencing methodologies utilizing fluorescence detection, as well as in other applications such as chromatin mapping. There are therefore various compounds comprising a fluorescent dye moiety and a nucleotide triphosphate moiety commercially available.

Affinity labels:

Other single molecule detection methods have availed of compounds having well studied affinity interactions with other molecules, such as receptor-ligand interactions.

Genome sequencing applications of the present invention may thus avail of established affinity labeling and detection methods. Other applications of the present invention my also benefit from the application of affinity labeling and detection methods.

Various compounds comprising a nucleotide triphosphate moiety and a small molecule affinity moiety are commercially available and suitable as substrates for DNA polymerases. said compounds have been used, in conjunction with DNA polymerases, to effect the affinity labeling of various polynucleotide molecules, and thus labeled polynucleotides are routinely subjected to manipulations comprising the formation of an affinity association with an appropriate receptor molecule. Two common examples are the use of biotin as said affinity moiety and streptavidin as said receptor molecule, and digoxigenin as said affinity moiety and anti-digoxigenin antibodies or fragments thereof as the respective said receptor molecule. It will be obvious to those skilled in the relevant arts that there are numerous other possible ligand-receptor interactions which may be exploited for affinity labeling purposes as well as immobilization purposes of the present invention, and that multiple distinct affinity interactions may be employed simultaneously.

For detection purposes, said affinity labels may be used to bind a microscopic colloid or bead which has been modified with an appropriate complementary affinity group such as a receptor.

Affinity label detection with microscopic beads:

In recent years a number of different methods and materials have been developed to permit the affinity binding of beads to molecules. Such binding is commonly accomplished by coating said beads with receptor molecules, such as streptavidin or Protein A (also known as Staph A, to which immunoglobulin G antibodies may subsequently be bound). Bead types include polymeric spheres of micron or submicron dimensions, metallic colloids such as colloidal gold, silica beads and magnetic beads. As will be obvious to those skilled in the art of polymer chemistry, polymer beads including dendrimers may incorporate dyes or liquid crystal molecules as side chains or within polymeric backbones, and these may facilitate optical detection methods. Attachment of appropriate receptor or affinity molecules to the surfaces of such beads yields a reagent suitable for the detection of an affinity labeled molecule. One such detection scheme was utilized by Finzi and Gelles,[23] albeit for different purposes.

Multimeric labels:

Where sensitivity to a single labeling moiety is insufficient, labeled reagents may comprise multiple occurrences of said labeling moiety in a manner that does not interfere with the corresponding molecular recognition and monomer addition processes, to increase the likelihood of correct signal amplification of any labeled molecule. For example, the ordinary single biotin moiety attached to a nucleotide by a linker may be replaced with a polymer having multiple biotin moieties as side chains, such that the likelihood of a streptavidin molecule interacting with this multimeric affinity label is increased. Fluorescent labels may similarly multiplied, as may any other labeling moieties. Measures must be taken in the design and synthesis of such multimerically labeled reagents to ensure that solubility is retained. This may be accomplished by choosing a highly soluble polymer as the backbone carrying said labeling moieties comprising the multimeric label.

Polymerization nucleating labels:

Any compound capable of serving as an initiator for some aqueous polymerization may also serve as a labeling moiety. This initiator nucleates the formation of a perceptible polymer attached to the sample molecule. Such a polymer, may, for example, comprise multiple fluorescent moieties, or simple effect a local change in transmitance of light or a local change of refractive index. After detection has been accomplished, said perceptible polymer is degraded or otherwise removed from the sample molecule. Such polymerizations may be self-limiting, as is the case for some dendrimeric polymers.

For this label detection methodology, polymerization is caused to occur in a step after the labeled nucleotide is added to the sample molecule, and must proceed via a chemistry that leaves the sample molecule in tact. Degradation or removal of said perceptible polymer must also leave the sample molecule in tact. Subject to the above stated limitation, any polymer and respective detection method may be employed.

Enzymatic labels and conjugates thereof:

Photochemical labeling:

Various methods have been developed for the photochemical labeling of molecules and especially biological macromolecules. These include detection of affinity labels such as biotin with conjugates of streptavidin and an appropriate enzyme capable of catalysing the formation of a chromophore from a chromophorigenic substrate, or capable of catalysing a

photon liberating chemical reaction, as with the enzyme luciferase. Such photochemical labeling methods will be readily applicable as detection methods for various embodiments of the present invention.

Note that multimeric affinity labels accessible for simultaneous association with multiple such enzymes will enable greater signal amplification, as will secondary enzyme amplification techniques and other techniques known within the molecular biological and microscopic arts.

Cleavable linkers:

Labeling moieties are favorably in communication with or coupled to nucleotides via a linker of sufficient length to ensure that the presence of said labeling moieties on said nucleotides will not interfere with the action of a polymerase enzyme on said nucleotides. Linkers will also necessarily be of some minimal length when stepping control is effected through the use of various preformed enzyme-nucleotide complexes (as described below). Once a nucleotide has been added by polymerization to (the daughter strand of) a sample molecule, and the accompanying label has been detected, proper detection and discrimination of subsequent nucleotides requires the elimination of said accompanying label. This may favorably be accomplished through the cleavage of said linkers which have been designed and synthesized to admit of cleavage by treatments which will not degrade or otherwise modify the relevant state or information content of sample molecules.

Cleavability may be provided for in a number of ways which will be obvious to those skilled in the arts of organic and synthetic chemistry. For example, said linker may include along its length one or more ester linkages, which will be susceptible to hydrolysis, which may be sufficiently mild for various ester functional groups. Amide linkages may similarly be employed. Linkages comprising disulfide bonds within their length have been developed to provide for cleavability[24]: reagents comprising such linkages are commercially available[25] and have been used to modify nucleotides[26] in a manner which may be conveniently reversed by treatment with mild reducing agents such as dithiothreitol. Cleavable linkages may be provided so as to minimize the portion of the linker which remains on the sample molecule. Because polymerases are relatively tolerant of linkers which may extend from various atoms of nucleotide molecules, it is not, however, critical that all of said linkers be cleaved away from the nucleotides incorporated into said sample molecules in the process of label removal.

Note that commercially available biotin derived nucleotides frequently contain, along the linker joining said biotin moiety to said nucleotide

moiety, one or more ester or amide bonds, which is susceptible to cleavage by various chemical treatments.

Note also that for linkers comprising appropriate bonds along their length, enzymatic cleavage may be performed.

Dissociative cleavage

Note that cleavage of a labeling moiety may also be effected by the disruption of some affinity interaction which effects the communication between said labeling moiety and the nucleotide moiety. In such cases, moieties joined by non-covalent associations may, for example, be dissociated by physical or chemical changes which do not necessarily cleave covalent bonds.

Photolabile linkers

Photocleavable moieties may also comprise an intermediate portion of linkers joining labeling moieties to nucleotide moieties, such that upon photocleavage of said photocleavable moieties, communication between the termini of said linker is disrupted and the label moiety is liberated from the nucleotide moiety. Because photodeprotection or photocleavage reactions[27] generally proceed quite rapidly, with appropriate detection and photoexcitation means, detection, label removal and nucleotide incorporation rates per sample molecule may approach the limit imposed by any particular polymerase enzyme and the processivity of said enzyme. Long linkers with photocleavable termini have been synthesized[28]

Thermolabile linkers

Similarly, compounds which thermally degrade into two or more portions may comprise an intermediate portion of such linkers, such that thermal cycling may be employed to effect linker cleavage. Thermostabile polymerases may be conveniently employed in embodiments availing thermolabile linkers.

Photomodification:

Single dye molecule photobleaching has been directly observed.[29] Fluorescent labels of nucleotides, particularly when only one or a small number of such moieties are used for labeling, may be neutralized by photobleaching, such that while some product of said fluorescent label may remain in communication with the sample molecule (e.g. the daughter strand of a polynucleotide being sequenced) it will no longer provide a signal sufficiently strong to interfere with the detection and discrimination of subsequently added labels.

Beyond photobleaching of fluorescent labels, affinity labels with appropriate photochemical properties may be subjected to photochemical modification rendering them inert to binding, generally subsequent to dissociation of the corresponding receptor by appropriate means.

Chemical neutralization/deactivation

For affinity labels, fluorescent labels or any other labeling moieties, chemical modification appropriate to the chemistries of said labels which effects a change or reduction in the detectable signal provided by said label may be availed to prevent interference of said labels with similar or distinct labels subsequently added to sample molecules or complexes thereof.

Labeling with activation and thermodynamic decay:

Compounds such as spirobenzopyran, which have labile, structurally and photochemically distinct but interconvertible isomers, may be used as labeling moieties. Here, an excited state of such a moiety may be used as a means of detection. After said detection has been successfully effected, chemical modification of one or another state of such interconvertible molecules may then neutralize it. Alternatively, activation may cause such a label to convert to some unstable but discernible state, which then irreversibly degrades according to characterizable kinetics. Such molecules must be chosen so as to remain in said discernible state for a sufficient time period to permit detection, but reliably degrade (to completion for a population of such molecules) within a practical time period.

Binding reaction inhibition detection methods:

Agents which specifically inhibit binding reactions may be identified rapidly through the detection of molecules, of a diverse library each molecular species of which is uniquely labeled, not bound by particles of some sample which may comprise many different species, in the presence of a test reagent, which is labeled, and permitted to associate with said sample (preferably during a preincubation step before the addition of said diverse library to said sample,) in analogy to blocking antibody assays. Results are compared to those obtained with an aliquot of said diverse library and another portion of the same said sample. Such an assay may be performed for increasing concentrations of said test reagent.

Enzymatically enforced associations at defined molecular sites:

Methods are provided to enforce highly specific associations and reactions, including molecular recognition processes, on individual sample

molecules or on populations and subpopulations of sample molecules. These are described for genome sequencing applications, but the methods included thereunder have broad applicability, including to any molecular affinity interaction.

Enzymatically enforced template directed copolymer addition at defined site:

Controlled comonomer addition:

Various methods may be used to accomplish the controlled addition of monomers, including nucleotides and especially labeled or protected nucleotides, to the daughter strand of a sample template molecule.

Rate control or accommodation:

Means of slowing the time required for the addition of a single nucleotide to a sample molecule will circumvent the requirement of stepping control. This will be particularly applicable for detection mechanisms not requiring separate manipulation steps (such as the separate association of beads to affinity labeled sample molecules). For example, the four nucleotides, each respectively labeled with unique, removable or neutralizable fluorescent labels, may be added to appropriately primed sample template molecules in the presence of polymerases, at low concentrations. Said concentrations must be sufficiently low that two nucleotides are not added to the sample molecule in less than the time required to accomplish the detection of the first such addition. Because all labels are present in the observation field, detection is accomplished through the observation of the reduction of the Brownian motion of a fluorescent moiety due to its addition to the sample molecule, in close analogy to the experiments of Finzi and Gelles[30], but it will be noted that the change in mobility is much larger in the present case. Alternatively, detection may be understood to depend on an increase in the net residence of some fluorescent moiety within a defined region or the occupancy of such a region, above the occupancy arising from the background of unbound labeled nucleotides.

Such detection is preferably conducted with a scanning excitation beam fluorescence confocal microscopic method as described above, or with a scanning detection light path, as also described above. Conditions (particularly nucleotide concentration) are chosen such that on average less than one labeled nucleotide will be present within the area illuminated by such a beam or thus observed, so that a light pulse of appropriate frequency passing through, for example, the pinhole which effects the scanning of the excitation beam, may be used to photobleach or

photocleave the fluorescent label from the sample molecule after it has been detected to have been added to the sample molecule, without the appreciable accumulation of incidentally unlabeled nucleotides. Alternatively, an SLM may be used to spatially control illumination of the sample by an appropriate frequency of light to effect photochemical unlabeling, and thus permit the simultaneous unlabeling of multiple sample molecules.

This method may be understood as concentration modulated control of the kinetics of polymerization processivity, which is used to facilitate direct observation of successive addition of individual (labeled) nucleotides, with controlled unlabeling. Scanning rate and other instrumentation dependent parameters will influence optimal conditions and concentrations. Thus, direct observation of the addition of comonomers is dynamically observed, and sequence information for the respective sample molecule may be reconstructed accordingly.

Stepping control by equilibrium means:

A simple method to effect adequate stepping control for sequencing applications of the present invention relies on equilibrium control. In this method, nucleotides (which are labeled) are limiting, and there is a relative excess of sample molecules. Exonuclease activity intrinsic to most polynucleotide polymerases is circumvented by the use of alpha-phosphorothioate nucleotides (which are appropriately labeled) which are resistant to such degradation, in this method. Other nucleotide derivatives or analogs suitable as substrates for polymerases and yielding exonuclease resistant polynucleotides may likewise be employed.

As an example of equilibrium controlled stepping, a thirty-three-fold excess of sample molecules relative to labeled complementary nucleotides per cycle may be chosen. Polymerase molecules are preferably provided in excess of sample template molecules. Each sample molecule has a three percent chance of undergoing a single nucleotide addition. Nucleotides are rapidly depleted. Any sample molecule which has undergone one nucleotide addition has a further three percent chance, or in total approximately a 0.1% chance of undergoing a second nucleotide addition. For a sequencing segment run of 20 bases per sample molecule, each segment will experience an error contribution of (20)(0.1%) or 2% from multiple additions within a cycle. Such erroneous segment data will be conspicuous when oversampling is performed due to the correspondingly low frequency with which it occurs. Alternatively, for tenfold excess of sample molecules with respect to labeled complementary nucleotides, there is a 1% chance per base of multiple additions to the same molecule, or, again for sequencing runs of

20 bases, a 20% chance that a segment experiences at least one duplicate addition event. For five-fold oversampling, the binomial distribution indicates that there is approximately a 94.2% chance that three or more segments including a particular base contain correct data regarding that base. Any specific individual data error is highly unlikely to occur more than once for fivefold oversampling. Note that in practice such calculations will also have to account for label amplification error and label detection error, but these error contributions should be susceptible to reduction to manageably low levels.

More generally, for a ratio x of nucleotide molecules to sample template molecules with a complementary base properly located relative to the primer, for x<1 there is a probability p equal to x that a particular sample molecule will experience the addition of at least one nucleotide, and a probability p^k that any sample molecule will experience at least k nucleotide additions within the same sequencing cycle. Multiple nucleotide additions to a sample molecule within the same sequencing cycle will result in erroneous sequence information being obtained from said sample molecule. The probability (d) of such a multiple incorporation error occurring within the sequence segment data obtained from a particular sample molecule in a sequencing run of n bases will be less than $2(n)(p^2)$. The net sequence information per sample molecule obtained per sequencing cycle will be x bases, and the net sequence information for a sample with N molecules will be (x)(N) bases, which will be large for large N. For example, with x=.03 and $N=3.3\times10^{10}$, there will be a net raw data accumulation of approximately 10^9 bases per cycle, which, with one-hundred-fold oversampling (i.e. due to each sequence being represented 100 times in the sample) will yield 10^7 bases of data per cycle; for a desired segment length of n=15 bases, n/x=(15)/(.03) or approximately 500 sequencing cycles will be required per run, and the run will yield 1.5×10^8 bases of information. For polymerase fidelity of 95% (an extremely low value chosen for purposes of illustration) there will be a 5% error rate (e) per base or a segment error rate of (n)(e)=75% per molecule, but the probability of two erroneous sequence segments having identical sequences will be $e^2(1-e)^{n-1}$ for segments with a single base error, which will be the most frequent error species. For this example, this yields a 0.12% frequency. Methods similar to those used to determine consensus sequences may thus be employed to obtain highly accurate data in spite of less than perfect polymerization fidelity. Thus, fidelity error components will be negligible compared to multiple base incorporation errors. For this example, multiple base incorporation error components will yield an error rate of less than $(2)(15)(.03)^2$ or about 3% per molecule. Again, oversampling will readily

detect such errors, which will occur identically for two molecules with only $d^2=(.03)^2$ or less than 0.1% probability, yielding a far lower error rate for oversampled data.

Stepping control by removable protecting groups:

Stepping control may favorably be applied to any polymerization process useful within the scope of the present invention, including both genome sequencing and affinity characterization applications.

Template directed polymerization depends on the processive addition of comonomers at the terminus of a growing daughter strand as specified by the respective complementary base of the parent template strand. Complementarity may be enforced through molecular recognition of said complementarity of protected analogs of said comonomers with the appropriate base of a template molecule, by the action upon such protected comonomers of appropriate polymerase enzymes.

Numerous monomers which may thus be added but do not provide an appropriate chemical functional group for subsequent elongation of the polynucleotide strand to which they have been enzymatically added are known within the relevant arts, and are generally referred to as chain terminators. Any such terminators which may be chemically or photochemically modified, particularly in a manner not disrupting the sample molecule, to a form which may support subsequent addition of comonomers in the usual manner, may be employed to effect controlled stepping of polymerization addition.

Removable protecting groups are particularly advantageous for the genome sequencing applications of the present invention because they may be utilized to permit and ensure that exactly one nucleotide is added to a sample molecule per sequencing cycle. This will permit an even greater rate of data accumulation than may be achieved by equilibrium control methods, with which only a fraction of the sample molecule population per cycle yields data.

Photoremovable protecting groups may be used to gain similar advantage but further permit controlled spatial localization of deprotection. Examples of such nucleosides have been prepared.[31] Because photodeprotection reactions[32] generally proceed rapidly, with appropriate detection and photoexcitation means, processivity and nucleotide incorporation rates per sample molecule may approach the limit imposed by any particular polymerase enzyme.

Nucleotide analogs comprising such removable protecting groups preferably further comprise labeling moieties. A particularly convenient category of such compounds comprises a labeling moiety or multimer thereof

in communication with the nucleotide moiety exclusively through said removable protecting group. For such compounds, removal of said removable protecting group will simultaneously effect removal of said labeling moiety. Simultaneous removal of both protecting moiety and labeling moiety will conveniently prepare a sample molecule for the next sequencing cycle in a single step.

Enzymological evidence concerning binding of 3' acetate esterified nucleotides and 5'-triphosphate-3'-(nucleoside-5'-monophosphate) to the triphosphate binding site of *E. coli* Polymerase I supports the acceptability of 3' modified nucleotides as substrates for this enzyme.[33] Such protecting groups should therefore be compatible with either naturally occurring or genetically modified polymerases.

Note that in other applications of the present invention, primers comprising a photodeprotectable 3' hydroxyl terminus (which may be synthesized by the polymerization of an appropriate 3' protected nucleotide onto the unprotected 3' hydroxyl of an oligo- or poly-nucleotide, for instance, by the action of terminal deoxynucleotidyl transferase) may provide for the selective polymerization of a polynucleotide moiety selectable by control over illumination of the appropriate region of the sample. A polynucleotide moiety to which such a primer is hybridized and then selectively deprotected may thus be subjected to amplification techniques such as PCR in a selectable manner. Such modified primers shall simply be referred to as photoactivatable primers.

The 3' deprotectable nucleotides employed in some variations on the present invention may also find other uses in molecular biology and biotechnology. They may be used as chain terminators in conventional enzymatic sequencing methods. If such manipulations are performed, any species terminating in a particular base may be extracted from the resolution medium (conventionally polyacrylamide gel), deprotected and then subjected to other manipulations requiring an active 3' hydroxyl group, such as ligation.

Enzyme adaptation to specific substrates:

The emergence of resistance to chain terminating nucleotide analogs by various viral polynucleotide polymerases suggests a convenient method for the *in vitro* evolution of polymerases capable of using reversibly 3' protected nucleotide analogs, or nucleotide analogs which otherwise serve as chain terminators which may be reactively modified to form an elongation competent molecule after incorporation into a polynucleotide. Further selection constraints may be concurrently or subsequently applied to fidelity, as the inclusion of non-sense condons in the coding region of a

dominant lethal protein coding gene which is carried by the same genetic material carrying the polymerase gene under selection, such that misreading of the non-sense codon, by the polymerase under selection, will effect lethality to the host and thus select against low-fidelity polymerases.

As stated above, such deprotectable compounds may serve as a convenient stepping control means for polymerization. Included among such deprotectable nucleotides are nucleotides with photocleavable protecting groups, including those which reside on the 3' hydroxyl of a nucleotide.

Label encoding and labeling methods for data collection:

Various systems may be used to represent the data corresponding to the occurrence of an affinity interaction. The complexity required of such a representational system will be determined by the types of molecules and associations being examined and the extent to which manipulative steps are to be minimized.

The most rudimentary encoding system will be a one-bit binary labeling system, consisting of only one label moiety type, indicating whether or not an association of only one resolvable type occurred during the preceding association step.

For example, consider a sequencing application employing only a single nucleotide labeling moiety. Such a system may avail each of the four nucleotides modified with a biotin moiety attached by a sufficiently long, cleavable linker arm. In such a case, a polymerization sub-cycle comprises: the incubation of sample template molecules bearing appropriate primers with an appropriate polymerase and limiting quantities of only one labeled nucleotide (and no unlabeled nucleotides) such that this monomer will be added only if the template molecule has the complementary base in the template position immediately 5' to the base opposite the 3' terminal base of the primer, and no monomers will be added otherwise; sample molecules are then washed to remove any remaining free nucleotides; the sample is then exposed to excess quantities of streptavidin modified fluorescently labeled beads for a sufficient length of time to ensure that all biotin moieties are bound by said labeled beads, and then all unbound beads are washed away; detection is then performed and data recorded; linkers are then cleaved. Said sub-cycle is repeated for the remaining three nucleotides, to constitute a cycle which successively tests for the presence in the sample template molecule of each type of base immediately 5' to the base opposite the 3' terminal base of the primer. If a sample molecule does not bind any label through such a cycle, then it was most likely "missed" due to the limiting concentration of nucleotides used to effect stepping of polymerization. If a sample molecule is labeled

multiply during such a cycle, then the respective subsequent bases are detected as occurring in the template according to the pattern of labeling.

A somewhat more efficient encoding system is provided if two distinct labeling moieties may be availed. Each nucleotide will be indicated by the presence or absence of each of the two moiety types, as a binary code. The moieties may, for instance be biotin (B) and digoxigenin (D). For example, the representation may be: A=B+D; T=B; G=D. These three nucleotides are added for a first polymerization sub-cycle, and all unbound reagents then washed away. Either two perceptibly distinct bead types may be used for simultaneous detection, provided distinct affinity labels are sufficiently well separated by extended linkers for simultaneous binding, or a single bead type with two distinct receptor molecules may be used in two separate binding and release cycles, in which case the release of one bead type will have to leave the remaining affinity moiety bound to sample molecules. After detection of bead labels, all remaining beads are removed and a second subcycle with C nucleotides affinity labeled with only one moiety are then polymerized onto sample molecules and appropriate detection is performed. Where protecting groups are used to effect stepping control, only one sub-cycle is needed and C may be unlabeled. In such cases unlabeled molecules will be detected as having added a cytidine.

More conveniently, nucleotides of each of the four types distinctly labeled with a fluorescent dye moiety may be used with fluorescence detection means, and a sequencing cycle consisting of only one sub-cycle. Alternatively, four antibodies (or four other appropriate receptor molecules or affinity reagents) which each bind each of the four distinct dye moieties may be bound to each of four perceptibly distinct beads. In another arrangement, nucleotides may each be labeled with some distinct combination of multiple dye moieties, again encoding a unique binary label.

Preferred Embodiments:

Polynucleotide sequencing:

Polynucleotide sequencing with random surface immobilization and light microscopic detection of affinity labels coupled to microscopic beads:

A DNA sample is prepared by shearing or digestion at a first sequence with a first restriction enzyme producing a 3' overhang terminus, to some appropriate, known size distribution, and labeled with a digoxigenin bearing nucleotide by the action of terminal deoxynucleotidyl transferase. After such digoxigenin labeling, said DNA sample is then subjected to random internal cleavage, for example by shearing so as to produce a

population of molecules with an average length half that produced in the previous sizing step, or digestion with a second restriction enzyme recognizing a distinct, second recognition sequence. Sample molecules of said sample are then bound at some convenient surface density to a transparent surface modified with a monolayer or a sub-monolayer density of anti-digoxigenin antibody. Said sample molecules, which will thus be bound to said transparent surface by the 3' termini of one strand, are then subjected to treatment by a 3' to 5' exonuclease, which will only act at the 3' terminus which does not bear the digoxigenin moiety due to the hindrance of this latter 3' terminus by its interaction with the surface, preferably not to completion of digestion of susceptible strands. Thus primed DNA sample template molecules bound to a transparent surface in an end-wise manner are prepared.

Using a single nucleotide labeling affinity moiety in a manner similar to the example provided for one-bit binary labeling systems, utilizing for example each of the four nucleotides derrivatized to effect communication of said nucleotides with a biotin moiety via a chemically cleavable linker, such as those described by S.W. Ruby et al.[34] polymerization directed by the template provided by each involved DNA sample template molecule is effected with an appropriate DNA polymerase lacking a 3' to 5' exonuclease activity, such as Sequenase 2.0,[35] with only one nucleotide type present during each polymerization step sub-cycle, at sufficiently low concentration to effect equilibrium controlled stepping. Polymerization reagents are then washed away, and may favorably be recycled after quantitation and readjustment of respective labeled nucleotide content. After each such polymerization sub-cycle step, which will add a biotin labeled nucleotide to only a fraction of those sample template molecules having only the base complementary to the nucleotide of said sub-cycle located immediately 5' to the base opposite the 3' terminal base of the strand priming this nucleotide addition, biotin bearing molecules may be labeled with microscopic streptavidin coated beads. Unbound beads are then washed away. Bead labeled molecules may then be observed by a video microscope, and the position of said bead labeled molecules within a sample may be recorded by image analysis of digital images thus obtained, in a manner similar to that used by Finzi and Gelles[36]. Dithiothreitol or other reagents capable of cleaving said linker holding said biotin in communication with said nucleotide incorporated during the previous polymerization sub-cycle are then used to treat sample molecules to cleave said linkers and thus release said biotin labeling moieties and the beads which have bound to them. A wash step is then performed to remove said beads. The extent of bead removal may be checked with another video

microscopy detection step if needed; and further cleavage treatment may be performed if decoupling was not adequate. The same subcycle (comprising polymerization, bead association, video microscopic examination, bead and label cleavage and removal by washing, and optionally a bead removal confirmation video microscopic examination step) is then repeated in succession for each of the three remaining nucleotide types, to complete a full base sequencing cycle (which as noted may yield information about more than one base location for some template molecules according to the sequence composition and the order of sub-cycles, and no information for other sample template molecules). Multiple said base sequence cycles are repeated until enough data have been accumulated relative to the total complexity of the initial DNA sample. Recorded data are then used to reconstruct sequence information for a segment of each sample template molecule, and segment sequence data are then aligned by appropriate computational algorithms.

Note that this embodiment avails only existing and generally available materials and devices, relies on relatively simple manipulations which are known to be highly reproducible according to their general use in the relevant fields, but due to the novel process of the present invention may yield genome sequence information far more rapidly and inexpensively than highly complex robotic instruments with sequencing methods utilizing electrophoretic separation.

Note that microscopic detection may be performed with a computer controlled stepable sample stage to effect the automated examination of large surface areas and hence very large numbers of sample molecules.

Alternatively, the transparent substrate providing the surface for immobilization may be that of a spooled film, which may be advanced at an appropriate rate before the objective of said video microscope of the present embodiment. Further, with such a spooled sample arrangement, said film may be circular, and continuously advanced through multiple video microscope apparata and wells effecting polymerization sub-cycles, all in appropriate order such that benefit of full pipelining of each step may be enjoyed. The construction of such instrumentation and rudimentary robotic actuation systems will be straightforward to those skilled in the relevant engineering arts.

Surface immobilization with single photon detection of plural fluorescent labels coupled to photodetachable 3'-hydroxyl protecting groups:

Sequence determination may additionally effected by the random immobilization at some appropriate density of appropriately prepared and

primed sample molecules on the surface of a transparent film, and stepwise polymerization with some appropriate polymerase, of all four nucleotides, all of which are protected at the 3'-hydroxyl with a photolabile (and hence photoremovable) protecting group in communication with labeling moieties which distinctly correspond to each nucleoside base type of the respective nucleotide. Label incorporation is detected, for example by the scanned beam light microscopic methods of the present invention, or with highly sensitive CCDs, and assigned to the spatial region occupied by a particular molecule. Said film is translated appropriately such that the full complexity of the sample may be examined after each polymerization cycle. Data are recorded electronically and according to the molecule for which they are obtained. Illumination of the sample with an appropriate frequency and intensity of light to effect 3'-hydroxy deprotection and hence also labeling moiety removal is performed, and a wash step is performed to remove freed label. Such polymerization, detection and deprotection cycles are repeated until the sample is sufficiently well characterized.

Random and non-random immobilization to optical detection array devices with optical labels:

Detection and classification of pathogens in clinical samples:

Methods of the present invention may be combined with the immobilization of highly diverse libraries of binding specificities with either encoding labels or phenogenocouples, which may therefore be characterized dynamically and related to any detected binding of particles of interest from a sample. Clinical samples are interacted with said libraries. All retained material is then interacted with some general label such as a polynucleotide binding dye (e.g. ethidium bromide, DAPI) or some chromophorigenic or photoemissive or labeled competitive inhibitor analog reagent detecting some metabolically fundamental reaction such as ATP hydrolysis, or the presence enzymes catalyzing said metabolically fundamental reaction. Pathogens containing polynucleotides or capable of said metabolically fundamental reaction may thus be detected.

The essential features of such a system are massively parallel screening for affinity interactions, generalized labeling methodology, and automated sample characterization. Because pathogen culturing is not required, and many types of highly specific information may be obtained in one assay procedure, without any previous knowledge of the state of the organism from

which said clinical sample was obtained, this represents the basis for extremely powerful diagnostic methods.

Note that various implementations may distribute binding specificities of known composition in a spatially controlled manner, and thus rely on spatial information to encode specificity type and hence, if known, composition of each specificity type. Note also that said libraries may comprise known mimetics or small molecules of known binding specificity.

The profile of any sample type from an individual organism according to such an assay may be monitored over time, and a profile is preferably obtained for a state of presumed health for comparison to samples correlated to states of disease, deficiency or degeneration or other states of ill health (i.e. longtitudinal tracking of individuals stratified by sample type). Samples of similar type may also be compared across populations and subpopulations, and the profile of these samples also correlated with state of health of the respective individuals (cross-sectional comparison).

For additional selectivity of detection, such a sample characterized as above may be further characterized according to the immunocharacterization method below.

Automated immunocharacterization and cyber-immune detection:

Such a system resembles that used for the detection and characterization of clinical samples, except that said highly libraries of binding specificities comprises a large number of immunoglobulin specificities. Libraries comprising immunoglobulin specificities may include such specificities in the form of immunoglobulins expressed on bacteriophages viruses, or in the form of the phenogenocouples of the present invention.

Banks comprising all of the specificities of a library may be maintained as monoclones, and upon detection of a pathogen in association with one or more binding specificity contained in some library, and the identification and/or characterization of said one or more binding specificity, an aliquot of the respective said monoclone, from one of said banks, may be provided to the organism. Such analysis and provision of one or more monoclones may be automated and controlled by algorithms.

Similar rapidity and broad characterization advantages are attained here as with the preceding method for the characterization of clinical samples.

Massively parallel enzymological assays:

In a manner similar to the preceding embodiments, several enzymes contained within some sample may be analyzed according to their binding probability, binding duration or dissociation rate, and conformational or

phosphorylation or other status. Such assays may favorably be performed by the methods of the present invention, with immobilized libraries which may include competitive inhibitors, and with pre- or post-binding labeling of sample enzymes by encoded label antibodies, to permit classification of sample enzyme type on a molecule by molecule basis, which classification data may be combined with the data obtained in this assay.

Additional Embodiments:

Hybridization based detection of polynucleotide sequences:

Various methods have been developed to test for the presence of short polynucleotide sequences and combinations of such sequences (according to stringency) in polynucleotide samples by hybridizing oligonucleotides or polynucleotides of known sequence to said polynucleotide samples. Such methods are sometimes termed "gene-probe" methods and often involve the use of immobilized, ordered arrays of oligonucleotides of known composition. Said ordered arrays have been formed on the surfaces of integrated electronic devices. It has been shown that, provided stringency can be made sufficiently high to prevent binding with even one base mismatch, such methods may be used to obtain sequence information about a sufficiently small sample.

The methods of the present invention provide a more rapid and convenient method for testing for the binding of known oligonucleotides to a complex polynucleotide sample, owing largely to the higher degree of parallelism which may be accomplished with single molecule methods. Here, each oligonucleotide, of known sequence, to be used as a specific gene probe, is synthesized with some perceptible encoded label, as described above, where the codes assigned to the sequence of said each oligonucleotide are known (due to the synthetic scheme by which they are produced and concurrently labeled). These are then hybridized to sample polynucleotide molecules, which either have previously been or will subsequently be immobilized, or may otherwise be separated from probe oligonucleotides, and the presence or absence of said each oligonucleotide in the sample polynucleotide containing fraction, which is a direct result of the success or failure of said each oligonucleotide to bind said sample polynucleotide molecules, will be readily ascertained through the detection and discrimination of the perceptible encoding labels corresponding to said each oligonucleotide. Contrary to the conventional gene-probe methodology, known probing molecules are generally unbound in this variation of the method as may be used with the present invention.

If the complexity of the polynucleotide sample is not too large, and the population made up of said oligonucleotides is sufficiently large and

complex, preferably exhaustively enumerating all possible oligonucleotides of the respective and sufficiently long length, and provided hybridization may be sufficiently stringent, which stringency is affected by a large number of known factors but also has sequence dependent components, information about the binding of said each oligonucleotide, which may be related to the respective known sequence and by Watson-Crick pairing rules to the respective sample polynucleotide sequence segment (or by identity with the strand complementary to the strand to which said each oligonucleotide has bound) may thus be obtained. As with other methods, alignment of such data may yield information about the sequence of the sample. The methods of the present invention further provide for the quantitation of such oligonucleotide hybridization by way of counting the number of times a particular perceptible encoded label is retained by a said polynucleotide sample, which may be availed both in the monitoring and correcting of errors and in the modulation of binding (hybridization) conditions.

Alternatively, probing may be accomplished by oligomeric sequences immobilized in some known configuration, for example by spatially patterned methods such as those of S.P.A. Fodor et al.[37] or by the lattices produced hierarchially by the method of N.C. Seeman noted above but comprising an ordered array (the order of which is predetermined by the incorporation or association of single stranded oligonucleotides or other single stranded termini of known sequence into or with modular components used to build up said lattices) of short single stranded regions of known sequence and preferably one free terminus (so as not to hinder conformational changes required for hybridization), but detected by the methods of the present invention, where sample polynucleotides are labeled with some appropriate discernible label, such as the dye YOYO-1, to facilitate the detection of their presence in association with each of said oligomeric sequences.

A yet further variation for effecting the spatially predetermined distribution of, for example and exhaustively enumerated population of single stranded oligonucleotides, may be effected by the used of the methods of N.C. Seeman to produce a uniform two dimensional lattice with a repeating pattern of short single stranded sequences with photoprotected termini, for example all of the 256 possible 4-mers. Such a lattice may have a periodicity substantially smaller than the wavelength of visible light. Said short single stranded sequences may be comprise some synthetic backbone so as to be resistant to enzymatic cleavage, which backbone preferably also is non-ionic (for example, of alkyl or beta-cyanoethyl derivation, peptide-nucleic-acid composition, or methylphosphonate composition) so as to denature from a complementary sequence only at

markedly elevated temperatures relative to ordinary oligonucleotides. Thus, a pattern of oligonucleotide complexity may be distributed in a predetermined manner below the resolution of light directed patterning. Light patterning techniques may then be availed to spatially direct the photodeprotection of said short single stranded sequences at lower resolution. Such light directed syntheses are preferably terminated with some comonomer which will prevent exonucleolytic degradation of said short single stranded sequences, or all of said short single stranded sequences are of a polarity opposite to that specified by the exonuclease to be subsequently used. By this combination of methods, patterning resolution is not limited by the properties of light, but may avail of the convenience of light directed patterning at lower resolutions. After a known distribution of all possible single stranded sequences of sufficient complexity has thus been produced, a denatured, labeled polynucleotide sample produced by extensive nick translation, with fluorescent labeled nucleotides, of a naturally occurring polynucleotide sample is hybridized to said lattice. Hybridized molecules are treated mildly with a single strand specific nuclease, followed by an exonuclease, to degrade or by the same process to free those regions which are not bound to the probing said short single stranded sequences. Label incorporated into the nick translation products of said polynucleotide sample is then detected and spatially mapped by the methods of the present invention, and binding is thus scored according to the known probing said short single stranded sequences. This method thus avails the molecular parallelism made possible by the molecular recognition, high density and high resolution detection methods availed with the present invention.

Note, finally, that higher density patterning than attainable by conventional light patterning methods may also be effected by scanning probe lithographic methods, such as the use of NFSOM lithography with photodeprotectable groups.

The foregoing descriptions of embodiments reveals the general nature of the present invention so that others skilled in the appropriate arts can, by applying current knowledge, readily adapt or modify these procedures and means to any of a vast number of applications and with any of a large number of possible implementations without departing from the essence of the present invention. Such adaptations and modifications are therefore comprehended within the meaning and range of equivalents of the disclosed procedures, means and embodiments. The embodiments disclosed herein are therefore provided for purposes of description and not limitation. It is

further understood that the terminology employed herein is similarly chosen for purposes of description and not limitation.

---

[1]Williams, N; Coleman, P.S.: 1982. *Journal of Biological Chemistry, 257(6)*:2834.
[2]Canard, B.; Sarfati, R.S.; 1994. *Gene, 148*:1.
[3]For example, see Betzig, E.: Chichester, R.J.; 1993. *Science. 262*:1422.
[4]Trautman, J.K.; Betzig, E.; et al.; 1994. *Nature. 369*:40.
[5]Basché, Th.; et al.; 1995. *Nature. 373*:132.
[6]Perkins, T.T.; et al.; 1994a. *Science, 264*:819.
[7]Perkins, T.T.; et al.; 1994b. *Science, 264*:822.
[8]Betzig, E.; Trautman, J.K.; 1992. *Science. 257*:189.
[9]For example, see Drake, B.; et al.; 1989. *Science, 243*:1586. Note that coloidal gold labels are readily observed on a conductive substrate.
[10]For a review, see Feenstra, R.M.; 1994. *Surface Science. 299-300(1-3)*:965.
[11]Garcia, R.; 1994. *Appl. Phys. Lett. 64(9)*:1162.
[12]Harvath, L.; 1994. Methods in Mol.Biol.; 34,(Immunocytochemical Method and Protocols); 337. See also
Nie,S; et al.; 1994. Science, 266:1018;
Eigen, M; Rigler, R; 1994. Proc. Natl. Acad. Sci., 91:5740
Eigen, M; 1993. Gene, 135:37.
[13]Fodor, S.P.A.; et al.; 1991. *Science. 251*:767.
[14] Regner, M.; et al.; 1993. *FEBS Letters. 336(3)*:452.
[15]Finzi, L.; Gelles, J.; 1995. *Science. 267*:378.
[16]Perkins, T.T.; et al.; 1994a.
[17]Perkins, T.T.; et al.; 1994b.
[18]Brenner, S.; Lerner, R.A.; 1992. *Proc. Natl. Acad. Sci.. 89*:5381.
[19]Brenner, S.; Lerner, R.A.; 1992.
[20]Perkins, T.T.; et al.; 1994a.
[21]Perkins, T.T.; et al.; 1994b.
[22]For example, see Scott, J.K. and Smith, G.P.; 1990. *Science. 249*:386.
[23]Finzi, L.; Gelles, J.; 1995.
[24]Ruby, S.W.; et al.; 1990. *Methods in Enzymology. 181*:97.
[25]for example, from Pierce Chemical Co.,'of Rockford, IL., U.S.A.
[26]Ruby, S.W.; et al.; 1990.
[27]Rajasekharan Pillai, V.N.; 1980. *Synthesis. 1980*:1.
[28]Rajasekharan Pillai, V.N.; 1980. See for example compounds 173 and 177 of this reference.
[29]Trautman, J.K.; Betzig, E.; et al.; 1994.
[30]Finzi, L.; Gelles, J.; 1995.
[31]Rajasekharan Pillai, V.N.; 1980. See in particular compounds 153 and 157 of this reference.
[32]Fodor, S.P.A.; et al.; 1991.
[33]Kornberg, A; 1980. *DNA Replication*. pp. 115-7.
[34]Ruby, S.W.; et al.; 1990.
[35]Supplied by United States Biochemicals.
[36]Finzi, L.; Gelles, J.; 1995.
[37]Fodor, S.P.A.; et al.; 1991.

Claims:

whereforth, I claim:

1. A process and method for the massively parallel characterization of molecular recognition and related phenomena comprising the steps of:

(a.) a step conferring repeatable unique identifiability to each first particle of a large population of first particles, which may be diverse;

(b.) the association of some group of single or plural perceptible or amplifiable labels, which may or may not be diverse, with some group of second particles, which may also be diverse, said perceptible or amplifiable labels either encoding information regarding the identity of each member of said group of second particles, or merely randomly associated each member of said group of second particles;

(c.) the exposure of said first particles of step (a) to the labeled or otherwise identifiable second particles of step (b), optionally in the presence of one or more enzymes or catalysts, and optionally followed by transport to remove any unassociated said second particles;

(d.) the detection, by appropriate methods corresponding to the various labeling methods applied to said first and said second particles, of the association or non-association of individual said second particles of step (b) with individual said first particles of step (a);

(e.) an optional step detecting or discerning the individual identity or identities of either or both of some or each of said first particles of step (a) and some or each of said second particles of step (b), to a desired degree of discrimination or specificity, by appropriate detection methods;

(f.) recording, generally by electronic means, the data collected in steps (d) and (e) in a manner which preserves any obtained information concerning the identities or classifications and state of association of all associated and unassociated individual said first particles and said second particles, or some desired subset of said obtained information;

(g.) optionally effecting the decoupling or neutralization of said labels of step (b) from the complexes formed by the association of said first particles of step (a) with said second particles of step (b), optionally including transport of any decoupled labels away from said complexes;

(h.) optionally detecting and recording the continued presence of any of said labels not successfully eliminated or neutralized in step (g) in a manner that specifically notes which of said complexes retain said labels and the identity of said labels;

in appropriate order which may or may not be in the precise sequence described in (a) through (h).

2. A method according to claim 1 comprising a procedure with at least one occurrence of step (a); where necessary for the preparation of reagents, one occurrence of step (b); and zero, one or more additional occurrences or repetitions of one or more of steps (b) through (h).

3. A method according to claim 2 where data recorded in any of steps (d) through (h) is used to predetermine the performance and variation of parameters of subsequent performance of one or more of steps (c) through (h).

4. Algorithmic means for controlling the effectuation of the method of claim 3.

5. Automated robotic means for performing the effectuation of the method of claim 4.

6. Algorithmic means for the predetermination of the probabilistic variation of the structural composition of a population of particles to be selected from by the methods of the present invention, or established methods of *in vitro* molecular evolution, in subsequent steps according to data obtained by the methods of the present invention.

7. A method for the determination of the sequence composition of a polynucleotide according to claim 2 where said first particles of step (a) are a polynucleotide sample, said second particles of step (b) or step (c) are labeled nucleotides, and polymerases are included as

enzymes of step (c), where said polymerases are one or more polymerases selected from the group consisting of: DNA dependent DNA polymerases, DNA dependent RNA polymerases, RNA dependent DNA polymerases also known as reverse transcriptase, and RNA dependent RNA polymerases also known as replicases.

8. A method according to claim 7 where said labeled nucleotides comprise one or more labeling moieties are selected from the group consisting of: fluorescent dye moieties; affinity label moieties; biotin; digoxigenin; fluoroescein, rhodamine; anthranylate and derivatives thereof; dinitrophenol.

9. A method according to claim 8 where said labeled nucleotides are labeled by labeling moieties which are in communication with the corresponding nucleotide moieties via a linker selected from the group consisting of: chemically cleavable linkers; physically cleavable linkers; thermolabile linkers; photolabile linkers.

10. A method according to claim 9 where said labeling moieties are affinity labels and where detection is performed by means of complementarily affinity labeled optically perceptible microscopic beads and appropriate optical detection means.

11. A method according to claim 8 where said labeled nucleotides comprise a removable 3' hydroxyl protecting group selected from the group consisting of: chemically removable protecting groups; thermolabile protecting groups; photoremovable protecting groups.

12 A method according to claim 11 where said labeled nucleotides comprise a labeling moiety in communication with the nucleotide moiety via said protecting group.

13. The process of claim 1, where said particles of step (a) comprise a polynucleotide sample and said particles of step (b) comprise short oligonucleotide means or equivalent means for probing the sequence of said polynucleotide sample.

14. The process of claim 13 where said oligonucleotide means is caries a removable protecting group on its 3' hydroxy terminal, the removal of

which will yield a 3' hydroxy group suitable as a substrate, when bound to a sample template molecule, for nucleotide addition by an appropriate polymerase enzyme, where said removable protecting group is selected from the group consisting of: photolabile protecting groups, thermolabile protecting groups, chemically removable protecting groups; thermolabile protecting groups.

15. A process according to claim 14, where said removable 3' hydroxy protecting group is a photolabile protecting group, whereby said photolabile protecting group is removed selectively by appropriate, spatially delimited illumination of said oligonucleotides which have been bound to said sample template molecules, and a polymerase and nucleotides are used to transcribe or otherwise copy portions of thus hybridized and selected sample template molecules, with optional characterization or release of the polynucleotide thus produced.

16. A group of compounds comprising two or more distinct segments, of monomeric or macromonomeric polymeric or macropolymeric composition, where at least two of said distinct segments are held in communication via cleavable linkers, and at least two of said distinct segments are of compositions providing suitable chemical means for controlled polymerizability.

17. A method comprising the synthesis of compounds from said group of compounds of claim 16 comprising the steps of:

(a.) forming one or more first covalent bonds between one or more of a first of said distinct segments within the structure of molecules of first said compounds of claim 16 and molecules of appropriate second compounds, at a sufficiently low concentration such that only one molecule of said first said compounds of claim 16 is thus combined a single molecule of said appropriate second compounds;

(b.) forming one or more second covalent bonds between one or more second said distinct segments of the structure deriving in step (a) from a molecule of said first compounds and one or more segments of the structure deriving in step (a) from the identical said single molecule of said appropriate second compounds of step (a), at a low concentration such that only intramolecular bond formation occurs;

(c.) optional cleavage by appropriate chemical or physical treatments or by photocleavage of one or more of said linkers deriving from the structure of the compounds of claim 16;

(d.) addition of a second said group of said compounds of claim 16 and repetition of steps (a) through (c);

(e.) a sufficient number of repetitions of the steps comprising step (d) to produce product compounds of desired size and complexity;

(f.) optional cleavage by appropriate chemical or physical treatments or by photocleavage or thermocleavage of any desired one or more of said linkers deriving from the structure of the compounds of claim 16 which remain.

18. Molecules produced according to claim 17 where said first distinct segments comprises polynucleotide moieties and where said second distinct segments comprise polypeptide moieties and the sequence composition of said polynucleotide moieties corresponds to the respective sequence composition of said polypeptide moieties on a molecule by molecule basis and according to predetermined coding.

19. Molecules according to claim 18 where said predetermined coding is the same as the universal code or variations thereupon with which genetic material of living organisms encode polypeptide and protein amino acid sequences.

20. A process for immobilizing and then determining the sequence of the polynucleotide moiety of single molecules of claim 17 availing single molecule examination methods.

21. The use of compounds produced by the process of claim 17 in any process of *in vitro* molecular evolution comprising selection according to one or more properties from the group consisting of: separation according to affinity properties; separation according to bond formation with or cleavage of bonds included within the structure of a linker which holds some affinity label or optical label in communication with the molecule under selection according to catalytic properties of the molecule under selection; or according to optical properties.

22. The use of compounds produced by the process of claim 17 with the procedures of *in vitro* molecular evolution availing of selection

according to data obtained through single molecule examination or single molecule characterization techniques.

23. The identification and characterization of molecules selected according to claim 22, subsequent a step immobilizing molecules of interest, particularly according to a method for the determination of some portion of the sequence of the polynucleotide moiety thereupon.

24. Method and means of light microscopy comprising the selection and delimitation of the optical paths passing through of some fraction of the normal microscopic visual field, by appropriate optical means, and the accumulation of data derived from said fraction of the normal visual field by sensitive photodetection means.

25. Method and means of light microscopy according to claim 24 where said appropriate optical means comprises one or more components selected from the group consisting of: one or more controllably translatable pinholes; one or more electronically controlled optoelectronic array devices; one or more spatial light modulators; one or more laser diode arrays, one or more light-emitting diode arrays.

26. Method and means of light microscopy according to claim 24 comprising the scanning of a beam of controlled frequency, which may be of nearly resolution limited cross-section, through a sample, and the observation of fluorescent emissions occurring in said sample with both controlled detection frequency selection means and highly sensitive photon detection means.

27. Method and means of light microscopy according to claim 26 where said beam of controlled frequency produced in a manner permitting the dynamically controlled variation of said frequency.

28. Method and means of light microscopy according to claim 26 where said controlled frequency selection means is capable of dynamic adjustment or variation.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/02342

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :Please See Extra Sheet.
US CL :435/6, 287.2, 288.7; 935/77, 87, 88
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6, 287.2, 288.7; 935/77, 87, 88

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	WO, A, 91/06678 (ROSS ET AL.) 16 May 1991, see entire disclosure.	13
Y	Proc. Natl. Acad. Sci., Volume 91, issued June 1994, M. Eigen et al., "Sorting single molecules: Application to diagnostics and evolutionary biotechnology", pages 5740-5747.	1-3,7-15
Y	Science, Volume 264, issued 06 May 1994, Perkins et al., "Relaxation of a Single DNA Molecule Observed by Optical Microscopy", pages 822-825.	1-3,7-15

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search	Date of mailing of the international search report
16 JULY 1996	02 AUG 1996
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer DAVID REDDING Telephone No. (703) 308-0651

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/02342

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	Japan J. Appl. Phys., Volume 33, No. 3A, issued March 1994, M. Ishikawa et al., "Single-Molecule Detection by Laser-Induced Fluorescence Technique with a Positive-Sensitive Photon-Counting Apparatus", pages 1571-1576.	1-3,7-15
Y	The Journal of Biological Chemistry, Volume 257, No.6, issued 25 March 1982, N. Williams et al., "Exploring the Adenine Nucleotide Binding Sites on Mitochondrial F1-ATPase with a New Photoaffinity Probe, 3'-o-(4-Benzoyl)benzoyl Adenosine 5'-Triphosphate", pages 2834-2841.	1-3,7-15

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/02342

A. CLASSIFICATION OF SUBJECT MATTER:
IPC (6):

C12Q 1/68; C12M 3/00; C12N 15/00